Emerging model species driven by transcriptomics

Dissertation

zur Erlangung des akademischen Grades doctor rerum naturalium

(Dr. rer. Nat.)

vorgelegt dem Rat der Biologisch-Pharmazeutischen Fakultät

der Friedrich-Schiller- Universität Jena

von Alexie Papanicolaou,

BSc Genetics, MRes Evolutionary Genetics

geboren am 14. July 1981

in Athen, Griechenland

Gutachter

1. Prof. Dr. David G. Heckel, Entomologie Abteilung, Max-Planck-Institut für chemische Ökologie, Jena D-07745, Deutschland; heckel@ice.mpg.de
2. Prof. Dr. Stefan Schuster, Friedrich-Schiller-University Jena, Faculty of Biology and Pharmacy, Department of Bioinformatics, Ernst-Abbe-Platz 2, Jena D-07743, Germany; Stefan.Schu@uni-jena.de
3. Prof. Dr. Anthony D. Long, Department of Ecology and Evolutionary Biology, University of California at Irvine, 321 Steinhaus Hall, Irvine CA 92697, USA.; tdlong@uci.edu

Tag der öffentlichen Verteidigung: 21. Juni 2011

Abstract

English

This work is focused on 'emerging model species', i.e. question-driven model species which have sufficient molecular resources to investigate a specific phenomenon in molecular biology, developmental biology, molecular ecology and evolution or related molecular fields. This thesis shows how transcriptomic data can be generated, analyzed, and used to investigate such phenomena of interest even in species lacking a reference genome. The initial ButterflyBase resource has proven to be useful to researchers of species without a reference genome but is limited to the Lepidoptera and supports only the older Sanger sequencing technologies. Thanks to Next Generation Sequencing, transcriptome sequencing is more cost effective but the bottleneck of transcriptomic projects is now the bioinformatic analysis and data mining/dissemination. Therefore, this work continues with presenting novel and innovative approaches which effectively overcome this bottleneck. The est2assembly software produces deeply annotated reference transcriptomes stored in the Chado database. The Drupal Bioinformatic Server Framework and genes4all provide species-neutral and an innovative approach in building standardized online databases and associated web services. All public insect mRNA data were analyzed with est2assembly and genes4all to produce the InsectaCentral. With InsectaCentral, a powerful resource is now available to assist molecular biology in any question-driven model insect species. The software presented here was developed according to specifications of the General Model Organism Database (GMOD) community. All software specifications are species-neutral and can be seamlessly deployed to assist any research community. Further through a case studies chapter, it becomes apparent that the transcriptomic approach is more cost-effective than a genomic approach and therefore sequence-driven evolutionary biology will benefit faster with this field.

German

In der Molekular-, Entwicklungs-, Evolutionsbiologie und verwandten Feldern werden Modellorganismen genutzt um (vereinfachte) Prozesse zu entschlüsseln auf denen biologische Phenomene aufbauen. Die vorliegende Arbeit beschäftigt sich mit Spezies die über ausreichende molekulare Resourcen verfügen. Diese tragen dann als neu aufkommende “emerging-model” bzw. “question-model” zur Klärung grundlegender Prozessabläufe bei. Im Verlauf der Dissertation wird gezeigt wie Transkriptomdaten generiert und analysiert werden mit dem Ziel die zu untersuchenden Vorgänge zu verstehen. Dies ist sogar in Spezies möglich die kein Referenzgenom vorzuweisen haben. Die ursprüngliche Resource “ButterflyBase” hat sich dabei als äusserst hilfreich erwiesen, ist jedoch limitiert auf die Ordnung Lepidoptera und unterstützt lediglich die klassische Sequenzierungstechnologie nach Sanger. Dank der neuen Technologie des “Next Generation Sequencing” wurde die Sequenzierung von Transkriptomen kosteneffektiver. Jedoch kommt es nun zu Engpässen in der bioinformatischen Analyse, dem Data-Mining und der Verteilung der Daten. Die vorliegende Arbeit stellt neue und innovative Methoden vor mit denen diese Enpässe effektiv behoben werden: die “est2assembly” Software generiert Referenztranskriptome mit “deep annotations” die im Chado Datenspeicher lagern. Das “Drupal Bioinformatic Framework” und die neue Bioinformatiksoftware “genes4all” liefern einen speziesneutralen und innovativen Ansatz um standardisierte Online-Datenbanken und damit verbundene Servicenetzwerke aufzubauen. Alle öffentlichen entomologischen mRNA-Daten wurden mit “est2assembly” und “genes4all” prozessiert um “InsectaCentral” ins Leben zu rufen. Mit “InsectaCentral” haben Wissenschaftler neuerdings Zugriff auf eine leistungsstarke Resource die bei der Erforschung molekularbiologischen Prozesse in jeder beliebigen “question-model”-Spezies (innerhalb der Insekten) behilflich ist. Die Software wurde gemäß den Vorgaben der “General Model Organism Database (GMOD)”-Gemeinschaft entwickelt. Besonders hervorzuheben ist, dass alle Softwareanwendungen speziesneutral sind und somit übergangslos von Wissenschaftlern ausserhalb des Feldes der Entomologie angewandt werden können. Anhand einer in dieser Arbeit vorgestellten Fallstudie wird erläutert, dass die Erforschung der Funktionsweise biologischer Prozesse mit Hilfe der Transkriptomforschung deutlich kosteneffiktiver ist als mit Hilfe der Genomforschung. Davon profitiert vor allem die sequenzorientierten Evolutionsbiologie.

Table of Contents

General Introduction

Thesis Overview

Researchers of biology are interested in finding out how biological processes work and how they have come to be, i.e. evolved. In our work, we make use of of scientific method (observation, hypothesis, experimentation, hypothesis re-formulation and back to experimentation) in order to reach this goal. To experiment with too many unknown variables leads to weak conclusions because we cannot know which variable had a causal link to the observed effect. For this reason, not only do we use controlled studies but we also use model systems to infer processes which occur in a larger part of the natural world. These model systems have traditionally been experimentally tractable organisms. In functional and biomedical biology, for example, much fundamental work was done using the budding yeast *Saccharomyces cerevisiae*, the fruitfly *Drosophila melanogaster*, the plant *Aradidopsis thaliana*, the nematode *Caenorhabditis elegans*, the frog *Xenopus laevis* and mouse *Mus musculus*. These model systems, all laboratory animals, have a large array of resources and they are highly tractable experimentally – albeit for different reasons. For that reason we call them model species and for that reason we expend most of our resources in improving the capability in these systems. It is commonly perceived that a non-model species is everything else. However, no researcher is using a non-model species for their research unless it was a model for some question. Indeed, the dove was the organism of choice used by Krebs to elucidate the glycolysis pathway. **What is certain, however, is that non-model species have scarce resources.** But is that perhaps a trend that is changing? **Is technology assisting us in generating resources with a fraction of the cost?** Are these resources sufficient to allow us to call a 'non-model' species now a model (if the word actually is still valid)? And once we do generate new resources, how can we actually make use of them to address specific biological questions? **This thesis is probing these questions by focusing on a particular subset of genomic resources**: one that a single research can now generate in a straightforward manner and one which, by using the content of this thesis, is going to be of use to the wider scientific community.

Before 2006, i.e. prior the start of the work presented herein, genomic resources were scarce for non-model eukaryotic species. Medium or large scale molecular biology was not considered standard practice for non-model insect researchers. Before I dive into the transcriptomics it ought to be pointed out that the word 'model' is one of the most ill-defined words in genomics. There ought to be a distinction between the traditional model (the experimentally tractable laboratory model for functional biology), a model due to sufficient resources allowing the investigation of a variety of biological phenomena (resource-rich model) and a model species because it is the most appropriate

organism to investigate a specific biological phenomenon thoroughly (question-model). As a community, we often think that question-models which don't have a genome are non-model species and once one is sequenced, they suddenly can became models. **This, and a number of other concepts introduced in this introductory section the thesis, hope to assist the reader in understanding the rest of this body of work.**

The drive behind this work was not to investigate a particular biological phenomenon or to just built a resource. The main aim was to understand what are the bottlenecks affecting question-models today and attempt, successfully I hope, to remove them. The long term goal which this work supports is how to assist question-model species to overcome any resource bottlenecks and became thus resource-models. This is an important point as arguments have been published pointing towards a change of focus away from question-models and into resource-rich species (Crawford 2001; Murray 2000). **Chapter 1 argues that such a shift of focus is not beneficial to the wider community, that question-model species can become resource models and suggest how.** Further, question-models can benefit the larger research community. Indeed, the work presented here has been one of the efforts to provide bioinformatic community with a non-biomedical resource-model species angle and specifically with the issue of what reference sequence can we use in order to conduct meaningful biological experiments. With Drs Beldade and McMillan we argue for the need of a reference sequence in the form of a genome sequence. But is a reference genome needed for every question-model? We point out how affordable multi-species transcriptome sequencing will become and that it can be utilized to build a reference.

Earlier work (Papanicolaou et al. 2005) showed how even modest studies can jump-start a species' molecular tools. The timeline of this work coincides with a technological breakthrough in the production of sequence data, the so called 2nd or Next Generation Sequencing technologies (NGS). NGS caused a revolution in the way we conduct research, both in non-model (e.g. the stickleback *Gasterosteus aculeatus*) and model species (e.g. *Drosophila* sp population genetics). By removing the sequence bottleneck, however, he have allowed for another to evolve: data analysis and dissemination or in other words, the bioinformatic bottleneck. Data analysis and dissemination is an integral part of these and future technological breakthroughs and must be addressed in order to make use of NGS beyond the standard protocols offered by the relevant companies to the biomedical community. **Chapter 2 is about the first complete framework for analyzing transcriptomic data to create reference transcriptomes.** This est2assembly software was published in BMC Bioinformatics in December 2009. As resource developers coming across a novel problem, we often have one of two solutions. First, we can create a solution most suited for the particular dataset at hand and focus our energy in ensuring that this particular dataset is well

analyzed. As we often build the system from scratch, we call these ad-hoc solutions. The second option is to attempt to build a general solution, one which can apply well to more than the particular dataset and focus our energy in ensuring that the solution can be integrated into a larger body of work. Because these integrated solutions tend to be taken up outside our research groups, they tend to have longer lifetimes at the cost of requiring longer development times. By reading the two annual special issues of Nucleic Acid Research (the Web Server and the Database issues) or even journals such as Bioinformatics and BMC Bioinformatics, we can detect that the majority of bioinformatic research has been of the first type but this situation has been shifting recently.

The timeline of this work also coincides with the formation of the first global bioinformatics consortium, GMOD. The acronym GMOD stands for Generic Model Organism Database but a redefinition of M for Myriad has been proposed to be more inclusive (see http://gmod.org). GMOD was first developed when there were a handful of resource-model organisms for functional biology and it appeared that obtaining the genomic sequence of an organism was a very expensive proposition, taking months or years to accomplish. These days, however, the number of resource models or near-models has increased thanks to NGS technologies. As well as being a community, GMOD is also a suite of inter-compatible software, it is made up of databases, applications, and so-called “adaptor” software that connects these components together. As a consortium driven by the database and genome consortia of the main traditional model species, the GMOD group focuses on the analysis and dissemination of genome data; the other types of data are treated as ancillary data points in connection to a reference sequence.

Chapters 3 and 5 will show how data dissemination has been solved using these two different approaches. ButterflyBase (Chapter 3), is a custom-built resource for the lepidopteran community, powered by a number of computational approaches developed or improved during this work in order to be able to fully annotate all Lepidoptera transcriptomes with the then available computing power. It was published in January 2008 by Nucleic Acid Research and has been cited 23 times since then (source: ISI Web of Knowledge accessed 03 October 2010). It was, however, not an integrated solution. InsectaCentral (Chapter 5) is complete rebuild, it utilizes the concepts presented in the introduction, it allows the assembly of NGS data and is a more stable and community-driven resource. The informatic engine, or 'framework', which drives InsectaCentral, and can drive any similar database, was developed as a module for the Drupal Content Management System. **Chapter 4 elaborates on the development of this GMOD Drupal Bioinformatic Server Framework (GMOD-DBSF) module.** Like InsectaCentral and est2assembly, GMOD-DBSF is now part of GMOD (http://gmod.org/gmod-dbsf, http://gmod.org/est2assembly, http://gmod.org/InsectaCentral).

Thanks to NGS, we can more cost-effectively create reference transcriptomes and this work has successfully bridged the bioinformatic gap in relation to transcriptomics. Reference transcriptomes can be used to answer specific biological questions if the appropriate bioinformatic tools for dissemination and analysis are provided. **Chapter 6 deals with the usage of reference transcriptomes for investigating specific biological questions.** The end-result of such bioinformatic experiments is usually i) a set of candidate sequences which need to be investigated with traditional hypothesis-driven molecular research; ii) a better understanding of experimental design and iii) a suite of tools which comply with the software-design criteria mentioned above and can be seamlessly utilized by other bioinformaticians.

Data rich, non-hypothesis driven research

Biologists from ecological fields have long been engaged with large amounts of raw data. Molecular biology, on the other hand, has only recently encountered this problem primarily thanks to the completion of the Human Genome Project (Venter et al. 2001; Lander et al. 2001) and then the inexpensive production of large-scale raw data. This so called -omic revolution has opened a non-hypothesis driven, exploratory approach in molecular biology (Collins et al. 2003). The -omics fields is ill-defined at best (Greenbaum et al. 2001), but a utilitarian definition would be a branch of biology which deals with large amounts of raw data of a certain type (creating thus an -ome): genomics deals with genomic DNA (genome); transcriptomics with mRNA derived data (transcriptome); proteomics with protein data (proteome); metabolomics with small molecules derived from a cell's metabolism (metabolome) etc to an often nonsensical degree (fortunately, ecology has not been renamed to ecologomics). A characteristic of the -omics is that it is usually not hypothesis-driven even though the mechanism of a specific biological phenomenon may be pursued: the main strategy in -omic experiments is to detect statistically significant patterns in Large-Scale (LS) experiments. For example, the transcriptional activity of cancerous and non-cancerous cells from a specific tissue may be investigated in a time-series, without prior knowledge what the pattern, if any, might be. Specific genes may be consistently differentially expressed in cancerous cells and therefore form a candidate cadre, derive a hypothesis and drive a subsequent hypothesis-driven experiment. Any experiment with large scale data poses, however, a challenge on how a scientist can analyze results (Bickel et al. in press). Especially in transcriptomics, statistical methods had to be developed to deal with e.g. the issue of multiple testing without being too conservative (Robinson & Oshlack 2010); data points with a large number of dimensions (e.g. tens of thousands in transcriptomics); a low sample size (repetition) and a non-trivial degree of noise derived from both the technical assay and the biology of the organism (e.g. differences in genetic background). To confound matters, no LS experiment is large enough: it can never capture a complete picture, it will inevitably be a sample of the whole population of data (Tukey in Bickel). Solutions do exist and the usual methodology to generate a hypothesis is derived from Artificial Intelligence (AI): first a visual inspection with histograms, boxplots, regression plots etc can detect overall patterns. Reducing the number of dimensions can be achieved with a Principle Component Analysis, cluster or network analysis and reveal hidden patterns (Slonim 2002). Statistical analyses, such as False Discovery Rate (FDR) (Benjamini & Hochberg 1995), avoid multiple testing issues without forcing an over-conservative Bonferroni correction yet ensure that the hypothesis has a limited number of variables. Finally, experimental design, usually driven by prior knowledge, modeling or data from pilot experiments, ensure that data gathered will be informative. For example

in transcriptomics, the number of biological and technical replicates must be sufficient to account for biological and technical variability in order to be able to detect variability in gene-expression due to the treatment under study. Indeed, machine learning approaches such as Support Vector machine learning (SVM) and AI concepts in general have offered an increasing number of tools to molecular biologists. As a result of this exploratory science, a novel type of biologist is being increasingly needed: bioinformaticians capable of analyzing LS-derived data and form hypotheses for subsequent experiments. Within the constantly changing environment of technology, bioinformatics, however, is no longer the narrow field it once was. These days it is seen in a supportive role for a variety of other fields which perform LS experiments. Systems biology is a relatively nascent field blending AI concepts and biology. It investigates a system's interactions that give rise to function or behavior. Systems biology is dependent on theoretical models and utilizes large amounts of data to identify candidates which can improve the model (Kitano 2002). Ecological and Evolutionary Functional Genomics (EEFG) is an evolution of molecular ecology into LS experiments. It utilizes genomics approaches to study adaptation of organisms to changing environments, genome evolution and population genetics, as well as the role of genomic evolution in the evolution of complex phenotypes (Mitchell-Olds et al. n.d.). Question-model species which are not necessarily resource-rich are used and often part of the work involves generating new resources. Unlike systems biology, concrete mathematical models are sparingly used. In order for bioinformatics to support these fields, it has been expanding the repertoire of expertise to include not only algorithmic biology but also artificial intelligence and information technology approaches.

Bioinformatics: an expanded field

The application of informatic methods in biology is not recent: even though the term bioinformatic is recent, the field is not recent and is in fact an experiment-driven science. For example, few wet-lab molecular biologists would consider that performing a BLAST analysis, a structural prediction or a multiple sequence alignment is actually an experiment. Nevertheless, it is an experiment and the result is nothing more than a hypothesis associated with a statistical significance. The overall trend is that as larger amounts of biological data are being generated, computer science approaches are increasingly integrated in biology . As a result, the field of bioinformatics widens as a single bioinformatician can no more be an expert in all bioinformatic fields than, say, an evolutionary biologist can be an expert in all approaches used in evolutionary biology. Traditionally, bioinformatics had an algorithmic focus (Durbin et al. 2002) due to the initial need to produce robust hypotheses in the face of increasing amounts of exploratory data. Even though many algorithmic methods had been taken up by biologists already, the data analysis from the Human Genome Project enticed a further innovation: the integration of non-algorithmic yet important

Information Technology (IT) and Artificial Intelligence (AI) concepts. Such concepts aim to clarify and enhance our ability to effectively acquire (i.e. mine) useful data (i.e. data-mining) and formulate testable hypotheses often expressed in a formal language (e.g. SBML, the Systems Biology Markup Language or UML, the Unified Modeling Language) and improve our ability to synthesize heterogeneous data (Sauer et al. 2007). As the work presented herein may be read by both molecular biologists and computational scientists, the following presents a short overview of important biological, AI and IT concepts used in this work such as Expressed Sequence Tags, Controlled Vocabularies (CVs), ontologies or relational databases and schemas.

Non-algorithmic concepts from Articifial Intelligence

In genomics and related fields (i.e. -omics), one of the most popular AI concepts is undoubtedly the Controlled Vocabulary (CV) and the associated ontologies; exemplified best perhaps by the work derived from the Gene Ontology Consortium (Ashburner et al. 2000). Controlled vocabularies enforce the rigid definition of terms (e.g. 'orthology' or 'gene') for a particular context; the structured relationship between such terms forms the context's ontology. In the context of phylogenetics, for example, I could define a 'gene' as an item with a known, unique, possibly mutable nature but unchanging identity which can be mapped on the branches of the phylogeny; in my system I could then define an 'informative' 'gene' to be one whose existence can be identified in every branch (i.e. no missing data) but its exact nature differs between branches (i.e. exhibits variation). In this case, every gene is either 'informative' or not, it cannot be both. Should the bulk of the phylogenetics community agree to the above definition then we would have a formal phylogenetics CV for the word gene and one of it properties. If the community constructed a CV for every term in phylogenetics and build a relationship graph connecting every connectible term, then we would have produced a phylogenetics ontology. It is of importance to note the contextual nature of an ontology: the word gene could be defined very differently in e.g. molecular biology or quantitative genetics. Indeed in the latter, 'gene' is often confounded with what a molecular biologist would define as a locus. Context is, therefore, an essential part of the system. In addition, even though they are rigid definitions, terms can be altered when breakthroughs occur: novel phylogenetic methods utilizing missing data could allow for genes which have an unknown nature in some parts of the phylogeny and thus changing my above ad-hoc definition of phylogenetically informative (Stamatakis & Alachiotis 2010). Further, a CV and its ontology is only unambiguous when the entire community which uses it subscribes to it.

The need for an ontology can be intuitive: it removes any confusion regarding the definition of a term but see (Lazebnik 2004) for an entertaining example of how the apoptosis community wasted decades of work due uncontrolled vocabularies and over-reductionist approaches. There are also underlying factors demanding the need for such a structured data model. Semantic integration is concerned with classifying both a source (e.g. gene) and target (e.g. informative), then choosing an appropriate relationship term (e.g. 'is' or 'is not') and finally joining these together to produce a semantic conceptualization which a computer can be aware of. In the above example, it would be inefficient to ask the computer ad-hoc to give you all the objects on a phylogenetic tree which 'shows some degree of variation but not too much to prevent an orthology statement' and is 'identified in all branches of the tree'. The efficient approach would be to have mapped all genes to

informative or non-informative (manually or via a software) and then ask it to produce all informative genes. Further, if the definition of informative is altered (as in the case of our missing data example above), the underlying engine retrieving the data needs not any correction: only the software which maps the term informative to a particular gene needs to be changed.

Transversing an ontology graph

To bioinformaticians, the appeal of ontologies is not the fact that we can have a formal dictionary but, due to its graph structure, we can transverse between terms using defined relationship terms (e.g. in the above example, 'is' or more formally is_a). In molecular biology, the most typical relationship terms are part_of, is_a and derived_from. In the Sequence Ontology (SO), for example, a mRNA kind_of transcript; transcript part_of gene; exon part_of transcript [Eilbeck]. Each term is explicitly defined within each ontology and synonyms may exist (is_a synonymous to kind_of). These definitions also explicitly state how properties defined for a CV term can or cannot be transferred to a linked term: in SO, part_of transitivity generally allows us to link exons with genes because part_of relationships link these two terms. This relationship breaks down however when the nature of the subject term is radically different from the object (e.g. leg part_of body) (Eilbeck et al. 2005). This forces general ontologies (such as SO) to develop multiple part_of terms such as component_part_of or member_part_of: the former is defined as having a transitive relationship, the latter does not; though both are hierarchical. These complications are of use to us: we have the tools to create a semantic mapping of real biological data into a database which allows simple queries to rapidly retrieve data of interest. For example, in ButterflyBase (Papanicolaou et al. 2008), which is not ontology-aware, each contig has an explicit link to a Single Nucleotide Polymorphism (SNP) and also to annotations transfered via BLAST. A direct search for any non-synonymous SNPs of the *Bombyx mori* homologue of the *Drosophila distalless* is not possible and must proceed iteratively: i) find all contigs similar to *distalless*; ii) find their SNPs; iii) fetch the SNP's annotation and iv) if non-synonymous, report it. Even in an automated user-interface system, this approach is computationally expensive. In an ontology-aware database, a query could be performed seamlessly like so: a SNP is part_of a codon, which is part_of a EST, which is part_of a gene: the gene has the BLAST annotation of *distalless* and the codon has an annotation which defines the classification of SNP. Further, multiple ontologies show the true appeal of ontologies: because *distalless* has been annotated with the Gene Ontology term 'proximal/distal pattern formation' (GO:0009954), it would be trivial to request all non-synonymous SNPs of GO:0009954 and compare them to, say, genes not involved in development (GO:0009954).

A point is reached, however, where mapping each annotation to the nearest neighbor term (e.g. non-

synonymous to SNP, SNP to codon, codon to EST etc) will create such inflation in the database that would render it so slow that it would be practically of no use. In such scenarios, specifications have to consider what queries is a user most likely to need. For example, we might not expect users to be interested in SNPs present on an individual EST which is homologous to a particular gene. InsectaCentral (this study) is highly ontology-aware (it actually completely depends on them) but chooses to map SNPs and their classification (and also their alleles) directly to an Open Reading Frame prediction which is derived_from a contig [see figure]. The same query as above can be performed but we have moderated the size of our database. The unlikely but ideal query mentioned above can still be completed (due to transitivity of EST part_of contig) but transverses through the contig and thus is slower.

Concepts from Information Technology

A main aim of IT is to efficiently compile and disseminate data without losing integrity. The more limited the resources of a particular LS experiment, the more integral the need to invest in IT concepts. ButterflyBase spawned from a species-specific project (Papanicolaou 2005) and therefore had not integrated many of the concepts on which the specifications of InsectaCentral (this study) depends on (see the InsectaCentral Chapter for a comparative discussion).

A relational database is derived from relational calculus and relational theory [Codd] and attempts to manage data via a collection of relationships. We define each relation set as a table where each row is a specific relation. The architecture of all tables in a database is a schema. Using pre-defined links (keys) between the tables we can transverse a graph linking multiple relations in a manner similar to transversing an ontology except that ontologies are usually directed and acyclic (the root node never joins with the terminal nodes) graphs. A relational database can be normalized at varying degrees which would be outside the scope of this work to detail. For the purposes of understanding InsectaCentral, however, normalization enforces stricter controls on data integrity and eliminates redundancy by increasing the complexity of the relations (i.e. by increasing the number of tables). The main cost is that more computational power is needed to reconstruct the full relationships. Additionally, the increased complexity makes the underlying schema more difficult to understand and use: the increased Shannon entropy (the information content of a message; Shannon 1948) means that every data point is more valuable and thus choosing not to transverse through it in the graph (e.g. due to data loss, limited computing power or the programmer not understanding it) results in an unacceptable loss of information. For example, the code GO:0009954 is of no use to anyone unless we can determine that i) it is in fact the GO term for development, ii) we can find out how GO defines development. It is not uncommon for resources to utilize more than one database

schema. As it will be shown in the est2assembly Chapter, InsectaCentral utilizes a highly normalized schema for data warehousing and a denormalized schema for driving GBrowse, a computationally intensive user-interface tool. Further, CVs, when appropriately applied, allow for a higher degree of database normalization and not only extend data integrity but also the ability to curate the warehoused data.

Regardless of the degree of normalization, any schema encodes information with a specific method. When a data warehouse wishes to share the data with another warehouse, the limiting step is to have a set of methodologies, a framework, to map data. Adopting an identical schema avoids this concern and the generalization of the FlyBase schema, Chado (Mungall & Emmert 2007), is a perhaps one of the most important IT innovations in molecular biology. Nevertheless, each data warehouse supports a different community and information is stored for different purposes. This is the first work which deals with information in emerging models and specifically with transcriptome resources. For that reason, InsectaCentral had to also produce a novel data querying framework. A framework in software engineering, is essentially a set of methods to abstract the technicalities of one engine (the Chado database schema in this case) and produce a user-interface (e.g. the InsectaCentral website) via a programming interface or middleware (also known as APIs – Application Programming Interface). Recent interest in the rapid generation of such interfaces led to the development of a specific type of API: the Content Management Systems (CMS). The main task of a CMS is to abstract the visualization interface by the deployment of common functions. For the work presented herein, I made use of the Drupal CMS, a community-supported software available at http://drupal.org. Drupal can be of especial interest to bioinformaticians because it has a powerful programming API utilizing PHP. Further, Drupal is open-source, licensed under the General Public License (GPL) 2. It has powerful CMS features and can handle database connections, user authentication, content permissions and the visualization interface. It is modular and straightforward to deploy: allowing for the development of modules which are activated and customized by the users acting as administrators. It is secure: allowing for authenticated and unauthenticated (guest) users with different permission settings. The former can belong to one or more user-groups (roles) which can have one or more special privileges (e.g. administrative rights). Special privileges are defined on a per-module basis and a user can, therefore, have administrative rights on a small portion of a complete Drupal website; for example adding new BLAST databases but not creating new users. The modularization and module cross-communication, allows for modules to securely extend the functionality of an online resource as a whole and extend the functionality of other modules while keeping code at a minimum. As part of this thesis and to assist future development of bioinformatic software using Drupal, I wrote a number of Drupal modules which formed the

informatic engine of InsectaCentral.

Defining a model species

Even though CV concepts are found in an ever increasing number of papers (e.g. at the time of writing, the Gene Ontology paper has been cited 4,658 times) basic definitions can still be confounded. Prior to whole genome sequencing, it was commonplace to consider that a model species is one where the bulk of research with a biomedical link was focused on (biomedical models) or a species with experimental tractability (i.e. experimental models). With the advent of genome sequencing any species which has a complete genome sequence, at least in eukaryotes, (e.g. in Lepidoptera the silkworm *Bombyx mori* , in Hemiptera the pea aphid *Acyrthosiphon pisum*) can be considered a genomic model species. This definition is now also tenuous: 2nd and 3rd generation sequencing technologies (see Appendix A) have allowed for an increasing number of genome projects. Such project were initially eukaryotes with small genome sizes (e.g. Protists, fungi) but now include large insect genomes (e.g. the Lepidoptera *Helicoverpa armigera* and *Heliconius melpomene*) or micro-organisms with large and complex genomes. We often need to use the word 'model' for systems concerned with merit and utility for a particular question, as well as the general degree of experimental tractability. In the context of genomics, a model species is considered one with a genome sequence but in the context of transcriptomes, not every species with a transcriptome assembly ought to be a model species. Indeed, with this trend, it is likely we will evolve towards an inflation of adjectives of 'model species', removing thus any utility from the phrase.

I will, therefore, utilize the word 'model' in a differing fashion. I am differentiating between a 'resource-model' and a 'question-model' species. The former is defined as a species which has a sufficient array of resources to be of wide-use: amenable to both forward and reverse genetics, high experimental tractability and a large repository of existing knowledge. It needs not, however, to be useful to every particular question. The question-model species is, on the other hand, one which is the best available system to investigate a particular biological phenomenon (e.g. the Krebs metabolic pathway was not dissected with *Saccharomyces cerevisiae* but with the dove: the latter's flight muscles allowed sufficient amounts of energy-consuming tissue to be extracted; Crawford 2001). It should be noted that resource-models, unlike question-models, are not contextual but fixed until major technical advances allow a new species to be considered a resource-model. Further, a question-model species needs to only have sufficient resources to fully dissect the biological phenomenon in question. In insects, an obvious example of a resource-model would be *Drosophila melanogaster*; but a model for investigating the molecular basis of asexuality, in the same phylogenetic order, could be *A. pisum* since *D. melanogaster* does not have an asexual life-cycle. In plants, *Arabidopsis thaliana* is undoubtely the most developed resource model but, regarding the

origin of dioecy, the *Silene* genus would be a more appropriate question-model species. More often than not, though, question-species do not have sufficient resources to approach the relevant questions from a molecular perspective.

Biological structures

If we wish to understand what is the molecular basis of certain biological phenomena, we have to first have an operational framework of how information is transferred. The central dogma of molecular biology (Crick 1970) is still operational today (with certain exceptions which seem to increase in number) and states that information is stored in the genomic DNA of an organism, termed genome. This can be in the form of actively transcribed sequence or not (formerlytermed 'junk-DNA'). The transcribed sections of the genome are called genes; they contain two types of sections: those which will be encoded to amino acids (triplets of DNA bases form codons) and those which will not. The former group into exons, the latter are either dispersed between exons (introns) or reside at the two ends of the molecule (the 5' and 3' being, respectively, the beginning and the end) and form the UnTranslated Region (UTR). Genes are transcribed from a DNA biopolymer to an RNA biopolymer via RNA polymerase in order to produce single stranded messenger RNA (mRNA). Initially, an exact copy of the DNA sequence is made (accounting for Uracil, the RNA-specific pyrimidine which replaces the DNA-specific thymine) and shortly thereafter the introns are spliced out to produce a mature mRNA. Active transcription, or also known as expression, is not constant in either a temporal or a spatial scale: different nuclei express different genes at different times. Molecular biology can record the collection of mRNA from a particular collection of cells (tissue) at a particular point of (developmental) time and produce a library. In most cases, each mRNA will be translated into a protein (each codon codes for an amino acid), modified if needed (post-translational modification) in order to be finally used by the cell or secreted to perform its function away from the producing cell. In other cases, regulatory mechanisms will prevent the mRNA from being translated. The information is, therefore, stored in the genomic DNA (gDNA), transported via the mRNA state to each cell's translational machinery (ribosomes) to produce the final information as a protein. As with mRNAs, the collection of proteins in a particular tissue and point of time can be recorded. The collection of all mRNA molecules encoded by an organism in every cell and time is called a transcriptome and the respective collection of proteins a proteome. Even though, within an organism, the genome is a static entity, the proteome and transcriptome differ between cells and different developmental time-points. At each stage there are a number regulatory and modification mechanisms which allow for the extensive variation needed to make an organism function despite a static genome.

Technologies for large-scale sequencing

Until recently, the only efficient sequencing technology was dye-termination capillary sequencing, now commonly referred as the Sanger method (Sanger et al. 1977). It produces 500 – 800 bp long sequences using fluorescently labeled nucleotides which when excited by a laser emit light at a different wavelength. As the DNA molecule passes through the detector, the nucleotides are excited in succession. New sequencing technologies have since evolved and the state of the art is summarized at (Delseny et al. n.d.) with an informative graph on the decreasing cost of sequencing. Briefly, the pyrosequencing method also records the addition of nucleotides via a coupled reaction during the incorporation of a new nucleotide on a strand which is being synthesised in real-time (Ronaghi 2001). The incorporation is specific because it complements the nucleotide present in the existing strand. This is also the basis of the 454 sequencing which uses an array to massively multiplex the reactions and the detections. First generation 454 sequencing was termed GS20 (ca 100-200 bp), with subsequent improvements called FLX (ca 300 bp) and FLX Titanium (circa 400 bp; also known as XLR) (Rothberg & Leamon 2008). The 454 was the first instrument of the so called Massively Parallel Sequencing (MPS) and produces currently one million sequences per run. A major disadvantage is that the length of a series of identical nucleotides (known as homopolymers) is unreliably estimated during incorporation. Officially it supports the incorporation of 8 nucleotides but in practice as few as 4 or 5 nucleotides may cause problems (Chevreux, pers. communication). The competing platform is Illumina (formerlycalled Solexa) with first generation instruments producing 35 bp and subsequent generations increasing to 50 bp, 75 bp and 110 bp until the release of the latest Hi-Seq instrument producing 200 bp. The Illumina approach uses a PCR amplification on the chip which allows for a very high density array and thus a large number of sequences per run (an order of magnitude more than 454). The ABI SOLiD system offered a computational intensive innovation: ligation and not synthesis of pairs of oligonucleotides which have a mixture of fluorochromes. The emission is read in all four channels simultaneously and subsequently can be reconstructed. Sequences of ca 50 bp are now available. The disadvantage is that sequence quality information differs substantially from other more traditional detection approaches. The more recent Helicos sequencing target single molecule sequencing and does not make use of any amplification step. It produces ca 35 bp of sequence and may be most useful in sequencing raw RNA without the need to generate a cDNA template (Ozsolak et al. 2009). A number of the so-called 3rd generation NGS technologies are expected. Pacific Biosciences promises Sanger length sequence reads but also a clever innovation: the switching off of the detection laser. The excitation energy laser seems to be one of the main reasons for limited sequence

quality and by switching it off and on again repeatedly, longer sequences can be obtained. This so-called 'strobed' sequencing allows an additional property: the production of linked sequences at a relatively known distance (in proportion to time). Other methods such as 454 and Illumina can produce paired-end libraries where two sequences of known distance are identified in order to replace the older fosmid or BAC-end sequencing of the Sanger days. Strobed sequencing on the other hand can produce multiple such islands of known sequence within a larger strand by simply switching on/off the laser multiple times. The disadvantage with this unpublished method is that it remains to be tested with real world data. Another unpublished method is Ion Torrent from the same inventor as 454. The innovation is that it uses semi-conductor technology to measure changes in pH as nucleotides are incorporated. Without a laser or an optical detection system, machines can be less expensive and can be built small, making it possible for the first time to have a benchtop Next Generation sequencer producing 100-200 bp of sequence in its first generation. It suffers, however, from the same issues as 454 regarding homopolymers. Another approach which does not use an optical detection system is the Oxford Nanopore technology which is also still not commercialized. In this ‘strand sequencing’ method, current through a protein nanopore is measured as a DNA polymer passes through that pore. Changes in this current are used to identify the DNA bases (http://www.nanoporetech.com). During this thesis, due to availability only the 454 (GS20, FLX and XLR) technology was explored for generating a reference. The Illumina technology was producing short 35 bp reads which are not applicable for reconstructing the reference sequence of a eukaryote de-novo. It was used, however, in the digital transcriptomics study where levels of expression between treatments were explored (cf. Case Studies Chapter).

Identifying the transcriptome

Each cell in an organism has, at any specific moment of time, certain genes actively transcribed, i.e. “switched-on”. The transcribed messages, mRNA, can be captured from a total RNA extraction using poly-A selection. This extract is unstable in solution and so a double stranded complementary DNA (cDNA) library is made. This library can then either be plated out and selected clones sequenced using Sanger sequencing or sequenced en-masse using a NGS technology. If only the end of the molecule is sequenced , then we only identify a tag from the whole molecule Expressed Sequence Tags (ESTs). If the entire molecule is sequenced because it has been sheared (e.g. in the Illumina protocol), then we produce an RNA-seq dataset. In reality, even non-sheared mRNA molecules are fragmented by chance so that through ESTs we can reconstruct the entire message (Figure 1). For most purposes reconstruction is possible by clustering the sequences and/or assembling them. Information on the utility of EST sequencing for EEFG, even before NGS, can be

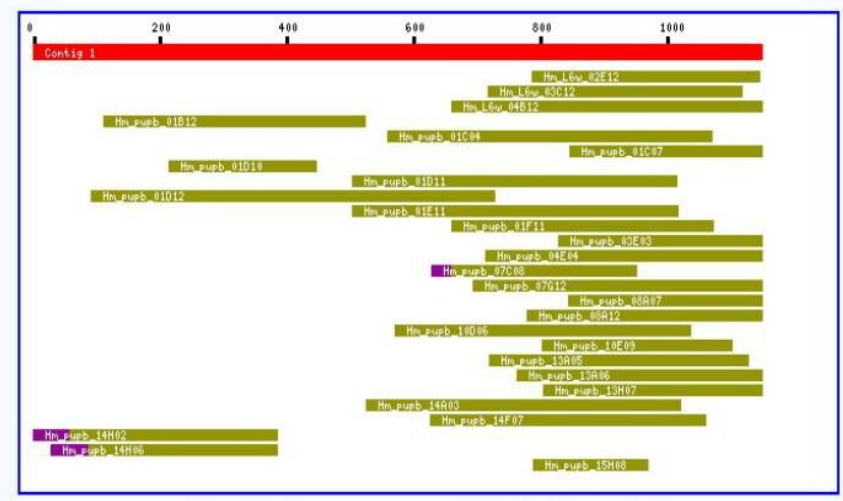


***Figure 1**: EST sequences from Heliconius melpomene (green) with vector sequence identified (purple). Because the ESTs are sequenced from random starting points, a clustering and assembly of the sequences allows us to reconstruct the hypothetical gene Contig 1 (red)*

read from (Bouck & Vision 2007). Briefly, they allow us to produce markers, identify genes without the need for a genome and learn about the transcription profile of individual cells. Each one of these methods has been enhanced with technology: shallow sequencing using Illumina allows us to produce thousands of markers in an inexpensive manner; microarrays allow us to survey the transcriptome of tissues quickly and efficiently even though new Illumina-based approaches (Chapter 6) are superceding the microarray approach. Most importantly, the long reads of 454 XLR sequencing allow us to sequence entire transcriptomes and identify all the genes of a species without the need for sequencing the genome first. This approach, which is still in its infancy, was the first target of this thesis. Further, researchers can choose to sample active genes of a specific tissue and/or developmental stage and thus increase their chances of detecting a specific group of genes. This approach can be used in a comparative fashion; detecting differences spatially, temporally, between species or environmental conditions.

Conclusion

For these data, however, to become useful to a wet lab biologist, they must be annotated and presented in a useful format. Subsequent chapters of this thesis deal with these issues. But can this transcriptomic approach support the transformation of resource-poor question-models into true model species? The brief answer is no but it is an essential first step. And a first step which can potentially reduce the resource bottleneck to a sufficient extend that new hypotheses can be formulated which could provide a new way of looking at a biological phenomenon.

Overview of the manuscripts

This thesis contains 5 manuscripts of which 4 are published (in 3 of which the candidate, A. Papanicolaou, is first author) and 1 manuscript is in preparation for submission:

Citations

- **Butterfly genomics eclosing**

  Beldade, P., McMillan, W.O. & Papanicolaou, A., 2007. Butterfly genomics eclosing. Heredity, 100(2), 150–157.

- **Next generation transcriptomes for next generation genomes using est2assembly.**

  Papanicolaou, A., R. Stierli, R.H. Ffrench-Constant, and D.G. Heckel. 2009. Next generation transcriptomes for next generation genomes using est2assembly. BMC Bioinformatics, 10(1), 447.

- **ButterflyBase: a platform for lepidopteran genomics**

  Papanicolaou, A., S. Gebauer-Jung, M. L. Blaxter, W. Owen McMillan, and C. D. Jiggins. 2008. ButterflyBase: a platform for lepidopteran genomics. Nucleic Acids Research, 36, D582-7.

- **The GMOD Drupal Bioinformatic Server Framework**
  Papanicolaou, A. Heckel, D.G. Bioinformatics (Oxford) 2010

  doi: 10.1093/bioinformatics/btq599

- **InsectaCentral: facilitating comparative genomics with one million insect proteins**

  Papanicolaou A., Heckel, D.G.. In preparation for DATABASE (Oxford University Press)

Outline & contributions of candidate

1. Manuscript 1 sets the theme of a thesis on bioinformatic bottlenecks for emerging model species. It argues that a shift of focus towards emerging models is useful to the wider community but question-model species must first become resource models by generating -omic resources and improving the bioinformatic bottleneck. The focus is on the *Bicyclus anynana* and *Heliconius melpomene* butterfly species but could be generalized. This perspectives paper was invited by T. Mitchell-Olds as part of a special issue on Ecological and Evolutionary Functional Genomics (EEFG). It was organized by P. Beldade and each author contributed an equal share in the drafting process. Section 1 (Butterflies as emerging model organisms in genomics; pages 151-153) were authored primarily by O.W. McMillan and P. Beldade; Section 2 (Genomic resources in butterflies; pages 153-155) by P. Beldade and A. Papanicolaou; Section 3 (Extending genomic research in butterflies; pages 155-156) by A. Papanicolaou and O.W. McMillan. A. Papanicolaou contributed circa 30 % of the total work (with O.W. McMillan 30 % and P. Beldade 40 %).

2. Manuscript 2 is about the first complete framework for analyzing transcriptomic data to create reference transcriptomes. Aimed at both large sequencing facilities (e.g. University of Edinburgh's Gene Pool service) and small genomic groups, it is responsible for reducing the bioinformatic bottleneck in reference transcriptome generation and standardizing the process. The est2assembly software is used by the InsectaCentral database presented in later chapters. The manuscript was organized by A. Papanicolaou. As indicated in the published paper: *A. Papanicolaou conceived, designed and performed the study; analyzed and interpreted data; coded the software and drafted the manuscript. R. Stierli co-authored the GFF writing software and the GBrowse schema. R. ffrench-Constant and D. G. Heckel drafted the manuscript, financed and provided infrastructure for the study.* A. Papanicolaou contributed to more than 90 % of the overall work.

3. Manuscript 3 provided the first dedicated transcriptome database which is widely used by the lepidopteran community (it was published in January 2008 and has been cited at least 23 times since then; source: ISI Web of Knowledge accessed 03 October 2010). The deployment made use of existing software (PartiGene) but the provision of reference sequence generated from transcriptome data for an entire taxon was innovative. This proof of concept paper showed how reference transcriptome research can benefit the EEFG field even when reference genomic sequence is lacking. The manuscript was organized by A. Papanicolaou. As indicated in the published paper: *The initial Heliconius EST database was*

conceived by C.D. Jiggins and M.R. Blaxter (and developed by A. Papanicolaou). The extension from the 'Heliconius ButterflyBase' to 'ButterflyBase' was conceived and developed by A. Papanicolaou with additional technical support from S. Gebauer-Jung. Intellectual support and motivation was from O.W. McMillan. This article was drafted by all authors. Further, A. Papanicolaou contributed to more than 90 % of the overall work.

4. Manuscript 4 produced the first bioinformatic library within the Drupal Content Management System (CMS). Included was i) a library for manipulating Chado and GMOD data (gmod-dbsf), ii) an innovative annotation server (biosoftware_bench) and iii) a module to database and disseminate RNAi experiments (genes4all_experiment). All are deployed within InsectaCentral and the latter was used in a recent review by Terenius et al (in press). The paper was organized by A. Papanicolaou and as indicated in the published paper: *A. Papanicolaou conceived, designed and programmed the software, co-ordinated and drafted the manuscript. D.G. Heckel tested the software, advised on design and drafted the manuscript.* Further, A. Papanicolaou contributed to more than 90 % of the overall work.
5. Manuscript 5 used the above manuscripts to build a unique database system for all Insects. Both the software and the database content are reported. The software is based on the FlyBase Chado database layout and uses the Drupal CMS to manage online content. It is build to be a robust, secure, easy to deploy and species-neutral solution so other laboratories can develop their own Central. The database contains all public insect transcriptome data (from Sanger and Next Generation Sequencing) and a number of secured pre-publication datasets contributed by collaborators. A. Papanicolaou conceived, designed and programmed the software, co-ordinated and drafted the manuscript. D.G. Heckel tested the software, advised on design and drafted the manuscript. A. Papanicolaou contributed to more than 90 % of the overall work.

Authorized by the dissertation supervisor,

Prof. David G. Heckel, PhD

Chapter 1 - Butterfly genomics eclosing

This Chapter sets the theme of a thesis on bioinformatic bottlenecks for emerging model species. It argues that a shift of focus towards emerging models is useful to the wider community but question-model species must first become resource models by generating -omic resources and improving the bioinformatic bottleneck. The focus is on the *Bicyclus anynana* and *Heliconius melpomene* butterfly species but could be generalized. This perspectives paper was invited by T. Mitchell-Olds as part of a special issue on Ecological and Evolutionary Functional Genomics (EEFG).

Citation

Beldade, P., McMillan, W.O. & Papanicolaou, A., 2007. Butterfly genomics eclosing. Heredity, 100(2), 150–157.

Heredity (2008) 100, 150–157

www.nature.com/hdy

SHORT REVIEW

Butterfly genomics eclosing

P Beldade[1], WO McMillan[2] and A Papanicolaou[3]
[1]*Section of Evolutionary Biology, Institute of Biology, Leiden University, Leiden, The Netherlands;* [2]*Department of Biology, University of Puerto Rico, San Juan, PR, Peurto Rico and* [3]*Department of Entomology, Max Planck Institute for Chemical Ecology, Jena, Germany*

Technological and conceptual advances of the last decade have led to an explosion of genomic data and the emergence of new research avenues. Evolutionary and ecological functional genomics, with its focus on the genes that affect ecological success and adaptation in natural populations, benefits immensely from a phylogenetically widespread sampling of biological patterns and processes. Among those organisms outside established model systems, butterflies offer exceptional opportunities for multidisciplinary research on the processes generating and maintaining variation in ecologically relevant traits. Here we highlight research on wing color pattern variation in two groups of Nymphalid butterflies, the African species *Bicyclus anynana* (subfamily Satyrinae) and species of the South American genus *Heliconius* (subfamily Heliconiinae), which are emerging as important systems for studying the nature and origins of functional diversity. Growing genomic resources including genomic and cDNA libraries, dense genetic maps, high-density gene arrays, and genetic transformation techniques are extending current gene mapping and expression profiling analysis and enabling the next generation of research questions linking genes, development, form, and fitness. Efforts to develop such resources in *Bicyclus* and *Heliconius* underscore the general challenges facing the larger research community and highlight the need for a community-wide effort to extend ongoing functional genomic research on butterflies.
Heredity (2008) **100,** 150–157; doi:10.1038/sj.hdy.6800934; published online 7 February 2007

Keywords: evolutionary and ecological functional genomics; butterfly wing patterns; *Bicyclus*; *Heliconius*; EST; linkage maps

Introduction

Genomics outside established model organisms

The initial lament that genomics 'would accelerate the migration of biologists to the 'superb six': humans, mice, fruitflies, worms, yeast, and *Arabidopsis*' (Murray, 2000) has failed to materialize (Crawford, 2001). Less than 5 years after the first draft of the human genome was published, nearly 600 eukaryotic genome-sequencing projects are completed or underway (cf. http://www.genomesonline.org/). The advantages of phylogenetically broad genome coverage are clear, and comparative analysis of diverse genomes will certainly continue to yield important insights into genome evolution and the relationships among branches of the tree of life. However, more than accumulating sequence data for comparative analysis, genomic research offers a unique opportunity to pursue a complete understanding of how genetic information is translated to produce an organism, and how modifications in genomic composition and organization give rise to biological diversity. In this quest, research on a new class of 'emerging' model organisms is an essential complement to the in-depth and finely detailed analysis of traditional genetic model organisms.

Correspondence: Dr P Beldade, Institute of Biology, University of Leiden, Kaiserstraat 63, 2311 GP Leiden, The Netherlands.
E-mail: pbeldade@biology.leidenuniv.nl
Received 13 April 2006; revised 16 October 2006; accepted 27 November 2006; published online 7 February 2007

Evolutionary and ecological functional genomics

The relatively new field of 'evolutionary and ecological functional genomics' (EEFG), and its goal of finding 'the genes that affect ecological success and evolutionary fitness in natural environments and populations' (Feder and Mitchell-Olds, 2003), requires an expansion outside classical model organisms. Model organisms for EEFG must combine broad genetic and ecological tractability with naturally occurring, functional variation (Feder and Mitchell-Olds, 2003).

Lepidoptera in general, and butterflies in particular, offer outstanding opportunities for integrative research at the interface between genomes and biological complexity. In spite of their immense biological (very species rich), economical (pests, pollinators and silk production), and societal (education and public understanding of science) value, available genomic resources in Lepidoptera have been limited. This situation is finally changing and independent efforts to develop core resources are underway for several species of butterflies and moths. Here we provide an overview of the strengths of butterflies as models in EEFG and summarize current efforts to develop resources in two groups, *Bicyclus* and *Heliconius*, which offer unique and complementary opportunities to study the links between genomic, developmental, and phenotypic diversity. Within this context, we discuss the general challenges facing the research community and highlight the need for a community-wide effort to consolidate and extend ongoing research.

Butterflies as emerging model organisms in genomics

The strength of butterflies as research targets derives from their extraordinary diversity, coupled with the exceptional opportunities to study the origins and maintenance of variation at nearly every biological level. The historical roots of butterfly research are deep, and the current research community is very active in a variety of areas of ecology and evolution (Boggs *et al*., 2003) ranging from the molecular details of insect color-vision (Briscoe and Chittka, 2001; Stavenga, 2002) to the analysis of human impact on biodiversity (Kotiaho *et al*., 2005; Mulder *et al*., 2005). Different species have provided some of the most important case studies on diverse topics in ecology and evolution. These include (1) population genetics and metapopulation dynamics focusing on the Glanville fritillary, *Melitaea cinxia* (Hanski, 2005), (2) long distance migration of the monarch butterfly, *Danaus plexippus* (Brower, 1996; Wassenaar and Hobson, 1998; Froy *et al*., 2003), (3) studies of Batesian mimicry, host plant detoxification, and pigment production in *Papilio* swallowtails (Li *et al*., 2003; Nijhout, 2003), and (4) evolution and development of wing patterns in the buckeye, *Junonia coenia* and a number of other species, including *Bicyclus anynana* and *Heliconius* (Beldade and Brakefield, 2002; McMillan *et al*., 2002; Marcus, 2005).

Evolution and development of butterfly wing patterns

Research on wing pattern formation is perhaps the most visually appealing example of the contribution butterflies can make to the understanding of the origins, maintenance, and modification of diversity. Virtually all of the more than 17 000 species of butterflies can be identified on the basis of the color patterns on their wings, and these highly diverse traits are emerging as invaluable systems for linking genes, gene networks, development, form, and function (Figure 1) (Nijhout, 1991; Beldade and Brakefield, 2002; McMillan *et al*., 2002; Brakefield *et al*., 2003; Evans and Marcus, 2006). Wings covered with colored scales (Figure 1d) are a morphological innovation of Lepidopterans and there is enormous pattern variation both within and across species. This variation is generally ecologically relevant and its adaptive value in natural populations has been extensively documented in relation to both biotic and abiotic factors (examples in Nijhout, 1991; Beldade and Brakefield, 2002; McMillan *et al*., 2002). Furthermore, the production and maintenance of this variation can be studied across a multitude of levels of biological organization (reviewed in Beldade and Brakefield, 2002;

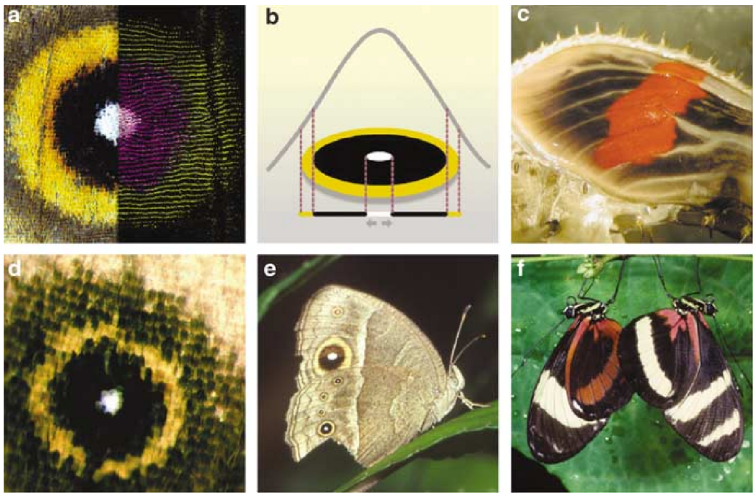


Figure 1 Multidisciplinary research in two colorful dimensions. Panels illustrate the different levels at which the mechanisms governing the production and modification of butterfly wing patterns can be studied. (**a**) A number of developmental candidate genes have been implicated in the formation of particular pattern elements such as eyespots. The genes *spalt* (pink) and *engrailed* (green) are expressed in pupal wings (right) in the center and in the different color rings of the future adult eyespot (left) (Brunetti *et al*., 2001). (**b**) Wing color pattern has also been studied in terms of the cellular interactions that underlie pattern formation and which are best understood for eyespots. In early pupal wings, the cells at the center of the presumptive eyespot produce a 'morphogen', which diffuses away from the center (arrows) to create a concentration gradient (gray curve). Neighboring cells then become fated to synthesize a particular color pigment depending on the morphogen concentrations they experience (where the vertical lines intersect the gray curve). (**c**) In *Heliconius*, the omochrome and melanin pathways (Nijhout, 1991) synthesize the pigment molecules that color the monochromatic scales. The deposition of different color pigments in different wing areas occurs late in pupal wing development (shown here for a pupa whose cuticle has been removed approximately 1 day before eclosion to expose the dorsal surface of the developing forewing). (**d**) The spatial arrangement of these scales in a single layer of parallel and overlapping rows produces the different pattern elements on the adult color phenotype (e.g. the eyespot on the photo). (**e**) Butterfly wing patterns play an important role in minimizing predation. The eyespots in *B. anynana*, for example, are thought to deflect predator's attention away from the fragile body as seen in this specimen photographed in the wild. (**f**) In addition, wing patterns have also been shown to play a role in mate selection and speciation. For example, the wing patterns in *Heliconius* provide both a source of ecological post-mating isolation and mating cues important in the incipient stages of speciation. See acknowledgements regarding source of photographs.

McMillan *et al*., 2002), ranging from the molecular details of pattern formation to the ecological relevance of pattern variation in natural populations (Figure 1).

During the last 15 years, explicit efforts to integrate methods and concepts from evolutionary and developmental biology have brought increased attention to research on butterfly wing patterns (Beldade and Brakefield, 2002; McMillan *et al*., 2002; Beldade *et al*., 2005; Joron *et al*., 2006a). This research has illustrated such exciting findings as the co-option of conserved pathways to produce evolutionary novelties (Brakefield *et al*., 1996; Brunetti *et al*., 2001; Reed and Serfas, 2004), the contribution of key development candidate genes to phenotypic variation (Beldade *et al*., 2002a; Kronforst *et al*., 2006), the mapping to the same genomic location of color pattern switch genes from different species (Joron *et al*., 2006b), and experimental tests of evolutionary constraints in morphological change (Beldade *et al*., 2002b; Frankino *et al*., 2005).

Two complementary systems

The African bush-brown *B. anynana* (Nymphalidae, Satyrinae) and species within the South-American genus *Heliconius* (Nymphalidae, Heliconiinae) have emerged as important players in research on how the reciprocal interactions between development and selection shape functional diversity (Figure 1).

The wings of these two Nymphalid clades are very different in shape and pattern (Figure 2). Furthermore, the striking phenotypic differences are accompanied by clear differences in ecological function and in the underlying genetic and developmental basis. Both groups are well suited for analysis at the molecular, organismal, and population levels and are textbook examples of natural polymorphisms. *Heliconius* is characterized by amazing geographic pattern divergence within species and pattern convergence between distantly related species (reviewed in Joron *et al*., 2006a), and *B. anynana* by striking seasonal variation and adaptive phenotypic plasticity (Brakefield and French, 1999). In both groups, wing patterns play a role in avoiding predation (Figure 1f) (Benson, 1972; Mallet and Barton, 1989; Kapan, 2001; Langham, 2004; Lyytinen *et al*., 2004; Brakefield and Frankino, 2006) and in mate selection (Figure 1e) (McMillan *et al*., 1997; Jiggins *et al*., 2001; Breuker and Brakefield, 2002; Robertson and Monteiro, 2005), but they seem to function in different manners. While the bright colors in *Heliconius* warn potential predators of the butterflies' distastefulness (Langham, 2004), those on *B. anynana* are associated to different seasonal strategies to avoid predation (camouflaging the dull-brown butterfly against a background of dry leaves, or attracting predators' attention away from the body against a green background; Figure 2a) (Brakefield and Frankino, 2006). These different ecological pressures lead to quite distinct modes of selection in natural populations: strong directional selection in *Heliconius* and divergent selection for opposite extreme phenotypes in the two seasonal environments experienced by *B. anynana* populations.

The genetic and developmental basis of wing pattern(s) formation also seems distinct in the two target groups. Study of laboratory populations of *B. anynana*

Figure 2 Extensive morphological variation in the wing patterns of *B. anynana* and *Heliconius* provide exciting opportunities for comparative work into the interplay between genes, development, and ecology. (**a**) Variation in *Bicyclus* wing patterns is extensive within and across species, and laboratory *B. anynana* provides the opportunity to study different types of variation (e.g. due to plasticity, to many alleles of small effect, or to single alleles of large effect) in detail. The *B. anynana* Stock Center in Leiden maintains over 20 lines with divergent phenotypes generated by artificial selection and over 30 mutant stocks carrying spontaneous mutations of large effect. The panel shows the ventral surface of both fore- and hindwing in different stocks of *B. anynana*. The first two photos on the left correspond to the 'wild-type' outbred stock and illustrate the seasonal polyphenism that results from plasticity in relation to temperature and humidity during development (Brakefield and Frankino, 2006) (on the left, a butterfly with conspicuous eyespots typical of the 'wet season', and to the right a dull-colored butterfly more typical of the 'dry season'). The remaining photos correspond to different mutant stocks with altered eyespot patterns (from left to right: *Bigeye* with enlarged eyespots, *spotty* with extra eyespots on the forewing, *Goldeneye* with the typically black ring replaced with golden scales, and *Missing* with two eyespots absent from the hindwing). (**b**) The radiation in *Heliconius* color patterns couples both divergent evolution and multiple independent cases of convergent evolution. Different *Heliconius* species can be easily maintained in captivity and different populations or closely related species can be crossed to study naturally occurring variation. The panel shows geographic variation in the mimetic species, *H. erato* (top row) and *H. melpomene* (second row). The two species fall on divergent lineages in the genus, yet share identical wing patterns across their sympatric ranges and have undergone a parallel radiation into over 30 different geographic forms (Sheppard *et al*., 1985). Color pattern variation in these species is largely explained by changes at 4–5 loci or complex of tightly linked loci of large effect. For example, allelic changes at the *Cr* locus in *H. erato* and in a complex of at least three tightly linked loci (*N, Yb, Sb*) in *H. melpomene* control most of the variation in yellow and white pattern elements among five geographic races shown. See acknowledgements regarding source of photographs.

have revealed both the presence of large amounts of segregating quantitative variation contributing to gradual response to artificial selection (Monteiro *et al*., 1994; Monteiro *et al*., 1997; Beldade *et al*., 2002b), and a number of spontaneous mutant alleles with a dramatic effect on phenotype (Beldade and Brakefield, 2002; Beldade *et al*., 2005). In *Heliconius*, in contrast, pattern variation is primarily attributable to a few genes of large effect with some minor effect modifiers (reviewed in Joron *et al*., 2006a). Differences in overall genetic architecture are emphasized by a more detailed analysis of specific candidate genes and pathways. The formation of butterfly eyespots, including those in *B. anynana*, involves expression of genes from classical wing development pathways (Brakefield *et al*., 1996; Brunetti *et al*., 2001; Reed and Serfas, 2004) in and around the area of the centers (foci) of presumptive eyespots with described organizing properties (French and Brakefield, 1995; Figure 1b). However, with the notable exception of tight linkage between *wingless* and the white/yellow color switch locus K in *H. cydno* (Kronforst *et al*., 2006), the bands and patches of color in *Heliconius* wings have so far shown no evidence for the involvement of the same developmental pathways (Reed and Gilbert, 2004; Jiggins *et al*., 2005; Tobler *et al*., 2005; Kapan *et al*., 2006; Joron *et al*., 2006a) or any type of patterning foci. Instead, genetic crosses and developmental mutants suggest that *Heliconius* patterns develop in a whole-wing proximodistal manner, independently of wing veins (Reed and Gilbert, 2004). These two seemingly distinct patterning systems within Nymphalid butterflies offer an excellent opportunity for a broad understanding of pattern formation and of the ecological consequences of variation in phenotype.

Genomic resources in butterflies

Advances in available genomic resources are fueling genome-wide research in *B. anynana* and *Heliconius*. The functional analysis of genotypic and phenotypic variants can be pursued both at the level of the molecular details of gene function during wing development (e.g. using spontaneous mutations and genetic transformation techniques (Lewis *et al*., 1999; Weatherbee *et al*., 1999; Marcus *et al*., 2004; Lewis and Brunetti, 2006)) and, at the other end of the spectrum, at the level of the ecological analysis of the adaptive value of variant phenotypes (Benson, 1972; Kapan, 2001; Langham, 2004; Mallet and Barton, 1989). Ultimately, this research promises to identify the genes and gene regions that underlie adaptive variation, link these to the genetic and biochemical networks responsible for pattern formation, and generate a fuller understanding of the interplay between genomic, developmental, and evolutionary processes.

Genetic sequence information

Construction and analysis of both cDNA and gDNA libraries is expanding the amount of sequence information available in butterflies. In the last couple of years, moderate-scale sequencing of expression sequence tags (ESTs) has catapulted gene discovery in *B. anynana* and *Heliconius erato* and *H. melpomene*. ESTs derived from developing wings (Papanicolaou *et al*., 2005; Beldade *et al*., 2006) have been independently assembled resulting in the identification of thousands of putative gene objects

Table 1 Resources in *B. anynana* and *Heliconius*

Species	*Genome size*	*Linkage map*[a]	*UniGenes*[b]
B. anynana	28 chromosomes 490 Mb 1361 cM[a]	352 AFLPs 8 msats	5721 (4251[c])
H. erato	21 chromosomes 395 Mb 1428 cM[a]	380 AFLPs 15 msats 16 SCNLs 9 isozymes	2981
H. melpomene	21 chromosomes 292 Mb 1616 cM[a]	219 AFLPs 23 msats 19 SCNLs	651

[a]Linkage maps for *B. anynana* (van't Hof *et al*., 2007), *H. erato* (Tobler *et al*., 2005; Kapan *et al*., 2006), and *H. melpomene* (Jiggins *et al*., 2005) include different types of markers: amplified fragment length polymorphisms (AFLPs), microsatellites (msats), single copy nuclear loci (SCNLs), and isozymes.
[b]Recent expression sequence tags (EST) projects have identified genes expressed in different tissues and developmental stages: *B. anynana* ESTs from five cDNA libraries made from developing wings at different stages, *H. erato* ESTs from a pooled cDNA library made from wing disc tissue collected from different geographic races at different developmental stages, and *H. melpomene* ESTs from three cDNA libraries made from whole-body pupae, late instar wing discs. All ESTs have been assembled and are available in *ButterflyBase*, www.butterflybase.org.
[c]The use of different assembly algorithms in *ButterflyBase* (Papanicolaou *et al*., 2005) and *openSputnik* (Beldade *et al*., 2006) explains the difference in total number of gene objects (*openSputnik* assembly in brackets).

(Table 1; Figure 3). These, together with publicly available sequences from other Lepidopteran species, have been assembled in a dedicated and web-accessible database, *ButterflyBase* (via http://www.butterflybase.org), designed to optimize the retrieval of individual ESTs or assembled gene objects annotated based on sequence similarity and protein prediction algorithms (Papanicolaou *et al*., 2005).

Gene discovery projects in *Heliconius* and *Bicyclus* have generated much sequence information, providing the first step towards enabling the study of genome evolution in butterflies. Many of the gene objects identified in initial EST scans showed similarity to genes in publicly available collections. This analysis has enabled the identification of genes from different functional categories, including genes known to be involved in insect wing development (candidate genes for wing pattern variation) and common 'house keeping genes' (valuable in comparative mapping studies, see below) (Papanicolaou *et al*., 2005; Beldade *et al*., 2006). However, there is a fairly large subset of coding regions that do not show clear homology to genes in publicly available collections (Beldade *et al*., 2006 and *ButterflyBase*), including those of the insect model *Drosophila melanogaster* and the Lepidopteran model *Bombyx mori* (with recently published genome (Mita *et al*., 2004; Xia *et al*., 2004) and large-scale EST projects (Mita *et al*., 2003; Cheng *et al*., 2004)) (Figure 3). Particularly exciting are a few hundred fairly large predicted peptides that may be new or highly diverged genes in butterflies (Papanicolaou *et al*., 2005; Beldade *et al*., 2006). A functional analysis of these genes (e.g. with analysis of patterns of gene expression) and the expansion of gene collections within butterflies will help to better characterize these emerging patterns. In this respect, the planned addition

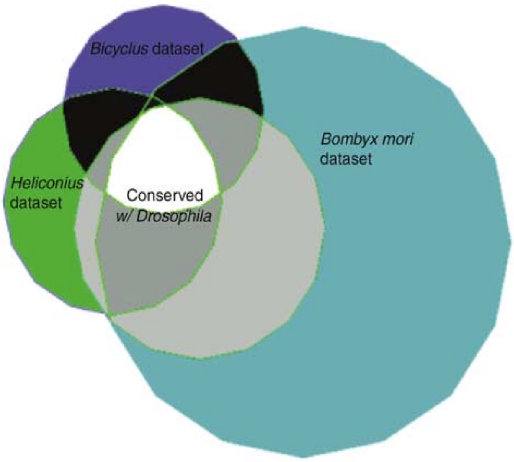


Figure 3 Overlap in EST-derived gene collections of *Heliconius*, *B. anynana*, and *B. mori*. The scaled Venn diagram (created using Vennmaster 0.17a; http://www.informatik.uni-ulm.de/ni/staff/HKestler/vennm/) shows the overlap between the collections of three Lepidoptera EST-cluster data sets from *ButterflyBase* (5721 clusters for *B. anynana*, 28 036 for *B. mori*, and 3632 for the pooled *H. erato* and *H. melpomene* collections) and the *Drosophila* proteins from *FlyBase* (69 920 peptide sequences from the Genome Annotation, Release 3). Each Lepidoptera collection was compared to the known *Drosophila* proteins using BLASTX similarity analysis, and to each other lepidopteran data set using BLASTN analysis. Lepidopteran gene clusters were assigned to the different areas of the Venn diagram based on a bit-score cutoff point of 70 bits. A total of 11 541 Lepidoptera clusters were significantly similar to proteins from the insect model *Drosophila*, and a subset of 1769 (white area) were conserved between all data sets. A total of 2161 gene clusters are shared across at least two Lepidoptera, but show no similarity with *Drosophila* peptides (black areas). The areas with clusters having no significant similarity to the other collections (in color) will likely decrease as the publicly available EST collections in lepidopterans increase, since a large proportion of these clusters likely reflect limitations of sampling cDNA (relatively few ESTs are available for butterflies) and sequencing (short reads make it harder to detect sequence homology). The use of a rather conservative estimate BLAST cutoff significance level (minimum 70 bits score corresponding to *e*-values lower than E−12) ensures lower rates of false positives (problematic when using gene collections that are not full sequence) but results in a potentially high number of false negatives (i.e. gene objects that do correspond to Lepidopteran homologs of annotated *Drosophila* peptides but which were not found significant here). Expansion of EST data sets for butterflies will enhance the estimates not only for large proteins but also of rapidly evolving genes or Lepidoptera- and Butterfly-specific genes.

of tens of thousands ESTs for *B. anynana* by the Joint Genome Institute (http://www.jgi.doe.gov) will provide an exciting data set of the genes expressed in different tissues and developmental stages in butterflies, and a powerful basis for comparative studies of Lepidopterans.

From identified genes to the genetic dissection of variation

The accumulation of sequence information is accelerating the development of the next generation of genomic resources in *Bicyclus* and *Heliconius* and expanding ongoing genetic mapping and expression profiling efforts.

High-density linkage maps predominately composed of amplified fragment length polymorphisms (AFLPs) and microsatellite markers are available for *B. anynana* and several *Heliconius* species (Table 1). These maps have been used to identify genomic regions that contribute to different types of phenotypic variation in the target Nymphalids (Jiggins *et al*., 2005; Tobler *et al*., 2005; Kapan *et al*., 2006; Joron *et al*., 2006b; van't Hof *et al*., 2007). Finer resolution mapping is being pursued by (1) adding gene-based markers throughout the genome (see below) and (2) by using linked AFLP markers and *bacterial artificial chromosome* (*BAC*) libraries now available in *H. erato*, *H. melpomene*, *H. numata*, and *B. anynana* to develop markers in genomic regions of interest. The latter strategy has been used successfully in *Heliconius* to show that the *NYbSb* gene complex in *H. melpomene*, the *P* locus in *H. numata*, and the *Cr* locus in *H. erato* all map to homologous regions of the genome (Joron *et al*., 2006b). This finding has been interpreted to suggest that a conserved, yet relatively unconstrained, mechanism affects pattern variation in *Heliconius*, and to imply that both convergent and divergent change can occur by the recruitment of homologous genomic regions. Positional cloning of these regions, now ongoing in all three species, will allow deeper insights into architecture, identity, and mode of action of this 'developmental hotspot' (cf. Richardson and Brakefield, 2003).

Current mapping efforts in both *Bicyclus* and *Heliconius* are concentrating on generating high-resolution gene-based maps. In this respect, ongoing EST projects are invaluable for the development of more markers for mapping and linkage analysis (Papanicolaou *et al*., 2005; Beldade *et al*., 2006). Sequence tags can be used to identify sequence polymorphisms in particular genes of interest, or, with targeted design, EST projects can directly combine gene and polymorphism discovery. In *B. anynana* and *Heliconius*, such a strategy has identified single-nucleotide polymorphisms and microsatellite repeats in thousands of gene objects (Beldade *et al*., 2006 and assembled ESTs in *ButterflyBase*). These types of markers are being added to existing linkage maps and will be a very powerful tool in moving from mapped regions to the identification of the actual genes that contribute to phenotypic variation. Particularly relevant are genes whose described role in wing development makes good candidates for wing pattern variation (cf. Beldade *et al*., 2002a). In addition, 'housekeeping' genes recurrent in EST projects of all species provide a common suite of reference markers for gene-based maps. Ribosomal protein genes, in particular, are ubiquitous in even moderate-scale EST scans and are excellent anchors for comparative linkage analysis (Yasukochi *et al*., 2006) (Table 2). Initial analysis based on ~30 orthologous markers mapped in *H. erato*, *H. melpomene*, and *B. mori* shows surprising levels of synteny (Jiggins *et al*., 2005; Kapan *et al*., 2006; Yasukochi *et al*., 2006). It will be exciting to confirm this observation for more markers in more species as the conservation of gene order would be a powerful tool to eventually identify mapped loci by comparison of maps from different species.

Gene mapping studies attempting to identify genes and gene regions contributing to variation in phenotype will be complemented with a detailed analysis of the changes in the levels of gene expression that accompany such variation. First generation high-density arrays

Table 2 Ribosomal proteins as candidate anchor loci for comparative mapping

Organism	*Taxon*[a]	*EST #*[b]	*RP #*[c]	% B. mori[d]
Heliconius melpomene	Papilionoidea, Nymphalidae	1258	91	88.3
H. erato	Papilionoidea, Nymphalidae	8129	73	70.8
Bicyclus anynana	Papilionoidea, Nymphalidae	9205	96	93.2
Papilio dardanus	Papilionoidea, Papilionidae	698	59	57.2
Euclidea glyphica	Noctuoidea, Noctuidae	570	16	15.5
Manduca sexta	Sphingoidea, Sphingidae	1991	83	80.5
Bombyx mori (B. mori)	Bombycoidea, Bombycidae	115 103	103	100.0
Plutella xylostella	Yponomeutoidea, Plutellidae	1129	91	88.3

[a]Classification within the order Lepidoptera (cf. National Center for Biotechnology Information (NCBI)).
[b]Total number of expression sequence tags (ESTs) in *ButterflyBase*.
[c]Number of ribosomal protein genes (RPs) identified in the different lepidopteran species using a BLASTN similarity search against the full coding sequences of *B. mori* RPs (GenBank AJ490511, AY578154, AY578155, AY583363, AY705974, AY706955-AY769343, DQ311196, DQ311216, DQ311285-DQ311290, DQ311379, DQ311405) with a cutoff threshold of 50 bits (cf. Parkinson *et al*., 2004).
[d]Number of ribosomal protein loci as a percentage of the total number of sequences searched against.

composed of genes expressed during wing development are being tested in both *Bicyclus* and *Heliconius* (Reed *et al*., 2007). These resources will allow expression profiling of different parts of the developing wing and different variants of the same species. Furthermore, the availability of BAC libraries will allow the characterization of regulatory regions around those genes whose map location or expression changes are associated with variation in phenotype. As the community continues to identify genes and genetic regulatory regions associated with pattern formation and pattern variation, the tools to test the functional importance of these loci are being perfected. Germline transformation technology has been developed in *B. anynana* (Marcus *et al*., 2004), and will be the basis for the next generation of functional experiments such as gene-targeted expression or knockouts.

Extending genomic research in butterflies

Core resources for genomic research in butterflies have expanded substantially over the last few years. However, for butterflies to fully emerge as ecological and evolutionary genomic models, commitment of the whole research community is required. A concerted effort is crucial to stimulate the development of shared resources and strategies are required to turn butterflies into competitive players in the genomics era and to enable a more complete analysis of the questions that have made this group such powerful biological models over the last couple hundred years.

Linking genomic, phenotypic, and ecological data

There is a rich history of collaborative multidisciplinary research in the butterfly community and the time has come to develop a common database containing both emerging genetic and genomic information and the vast amount of non-genomic data available for butterflies. Such database would link genomic/genetic diversity data (physical/linkage maps, expression data, ESTs, sequence polymorphisms, and genomic sequences) and phenotypic diversity data (quantitative and qualitative descriptions of phenotypes, images, and pedigrees) within the context of clear spatial (e.g. habitats and sampling sites) and temporal scales. Equally important is the development of common tools to utilize such a database and permit detailed queries across species collections. These are challenging issues that require broad community participation. Fortunately, we are not alone and the challenges faced by the butterfly community are identical to those faced by other emerging model groups including *Mimulus*, Cichlids, Sticklebacks, *Daphnia*, and *Dictyostelium*. A number of bioinformatics solutions to these challenges are available including, for example, GMOD (http://www.gmod.org), a generalized open-source resource fully equipped with standard ontologies, file formats, web site and database options, and tools for organizing genomic data.

Prioritize genome sequencing

Very importantly, the community must push forward efforts to get at least one butterfly genome sequenced. Genome sequence information will provide an invaluable anchor for all genetic and genomic research in this group. Genome projects in Lepidoptera are so far restricted to moths, with *B. mori* being the only published effort (Mita *et al*., 2004; Xia *et al*., 2004). It is hoped that newly available physical maps (Yamamoto *et al*., 2006; Yasukochi *et al*., 2006) will accelerate assembly and annotation of the silkmoth genome, but it is still unclear how far this resource can be used in a detailed genetic analysis of butterflies. Butterflies and the Lepidopteran lineage containing *B. mori* have probably diverged more than hundred Mya (Vane-Wright, 2004), and have quite distinct biological properties related to the contrast between the diurnal (in butterflies) and the nocturnal (moths) lifestyles. Unfortunately, the same diversity that makes butterflies such attractive models has so far made community cohesion challenging. While genome projects continue to be a major financial and technical undertaking, the community will need to rally behind one or perhaps two species to be able to make the strongest possible argument for sequencing a butterfly genome. The creation of a 'Butterfly Consortium', similar to what has been put together for other organisms, is necessary to fuel discussions and overcome these types of challenges. As new technology reduces the cost of sequencing and enables the addition of new genomes (see Bonetta, 2006), the community will be well positioned to capitalize on the strength of lepidopteran diversity to study a wide array of biological processes.

Butterfly genomics eclosing

These are exciting times, as we witness the metamorphosing of butterflies from classical organisms in ecological and evolutionary analysis to players in the genomics era. Indeed, research on *B. anynana* and *Heliconius* highlights the utility of butterflies as models for evolutionary and ecological genomic research, both satisfying essential EEFG criteria (Feder and Mitchell-Olds, 2003). With expanding genomic resources, EEFG on butterflies promises to provide important insights into the links between developmental diversity, phenotypic variation, and macroevolution. Ultimately, the combination of new tools, extraordinary diversity, and a rich history of research in ecology and evolution will ensure that butterflies can fully realize the long promised potential illustrated by the words of the nineteen century naturalist H.W. Bates, *'the study of butterflies – creatures selected as the types of airiness and frivolity…, will some day be valued as one of the most important branches of the Biological Sciences' (Henry Walter Bates, The Naturalist on the River Amazons, 1864).*

Acknowledgements

We thank our colleagues who shared images included in our figures: Craig Brunetti (panel 1a), Marcel Dix (1b), Bob Reed (1c), Paul Brakefield (1e), Larry Gilbert (1f), Antónia Monteiro (2a), and Jim Mallet (2b). PB is supported by the Dutch Research Organization NWO, WOM by the American NSF and the National Evolutionary Synthesis Center, and AP by the Max-Planck-Gesellschaft.

References

Beldade P, Brakefield PM (2002). The genetics and evo-devo of butterfly wing patterns. *Nat Rev Genet* **3**: 442–452.

Beldade P, Brakefield PM, Long AD (2002a). Contribution of *Distal-less* to quantitative variation in butterfly eyespots. *Nature* **415**: 315–318.

Beldade P, Brakefield PM, Long AD (2005). Generating phenotypic variation: prospects from 'evo-devo' research on *Bicyclus anynana* wing patterns. *Evol Dev* **7**: 101–107.

Beldade P, Koops K, Brakefield PM (2002b). Developmental constraints versus flexibility in morphological evolution. *Nature* **416**: 844–847.

Beldade P, Rudd S, Gruber JD, Long AD (2006). A wing expressed sequence tag resource for *Bicyclus anynana* butterflies, an evo-devo model. *BMC Genomics* **7**: 130.

Benson WW (1972). Natural selection for Müllerian mimicry in *Heliconius erato* in Costa Rica. *Science* **176**: 936–939.

Boggs CL, Watt WB, Ehrlich PR (eds) (2003). *Butterflies: Ecology and Evolution Taking Flight*. The University of Chicago Press. p 756.

Bonetta L (2006). Genome sequencing in the fast lane. *Nat Methods* **3**: 141–147.

Brakefield PM, Frankino WA (2006). Polyphenisms in Lepidoptera: multidisciplinary approaches to studies of evolution. In: Ananthakrishnan TN, Whitman DW (eds). *Phenotypic Plasticity in Insects*. Oxford University Press: Oxford.

Brakefield PM, French V (1999). Butterfly wings: the evolution of development of colour patterns. *BioEssays* **21**: 391–401.

Brakefield PM, French V, Zwaan BJ (2003). Development and the genetics of evolutionary change within insect species. *Ann Rev Ecol Evol System* **34**: 633–660.

Brakefield PM, Gates J, Keys D, Kesbeke F, Wijngaarden PJ, Monteiro A *et al.* (1996). Development, plasticity and evolution of butterfly wing patterns. *Nature* **384**: 236–242.

Breuker CJ, Brakefield PM (2002). Female choice depends on size but not symmetry of dorsal eyespots in the butterfly *Bicyclus anynana*. *Proc R Soc Lond Ser B-Biol Sci* **269**: 1233–1239.

Briscoe AD, Chittka L (2001). The evolution of color vision in insects. *Annu Rev Entomol* **46**: 471–510.

Brower LP (1996). Monarch butterfly orientation: missing pieces of a magnificent puzzle. *J Exp Biol* **199**: 93–103.

Brunetti CR, Selegue JE, Monteiro A, French V, Brakefield PM, Carroll SB (2001). The generation and diversification of butterfly eyespot color patterns. *Curr Biol* **11**: 1578–1585.

Cheng TC, Xia QY, Qian JF, Liu C, Lin Y, Zha XF *et al.* (2004). Mining single nucleotide polymorphisms from EST data of silkworm, *Bombyx mori*, inbred strain Dazao. *Insect Biochem Mol Biol* **34**: 523–530.

Crawford DL (2001). Functional genomics does not have to be limited to a few select organisms. *Genome Biol* **2**: INTERACTIONS1001.

Evans TM, Marcus JM (2006). A simulation study of the genetic regulatory hierarchy for butterfly eyespot focus determination. *Evol Dev* **8**: 273–283.

Feder ME, Mitchell-Olds T (2003). Evolutionary and ecological functional genomics. *Nat Rev Genet* **4**: 649–655.

Frankino WA, Zwaan BJ, Stern DL, Brakefield PM (2005). Natural selection and developmental constraints in the evolution of allometries. *Science* **307**: 718–720.

French V, Brakefield PM (1995). Eyespot development on butterfly wings: the focal signal. *Dev Biol* **168**: 112–123.

Froy O, Gotter AL, Casselman AL, Reppert SM (2003). Illuminating the circadian clock in monarch butterfly migration. *Science* **300**: 1303–1305.

Hanski I (2005). Landscape fragmentation, biodiversity loss and the societal response. The long term consequences of our use of natural resources may be surprising and unpleasant. *EMBO Rep* **6**: 388–392.

Jiggins CD, Mavarez J, Beltran M, McMillan WO, Johnston JS, Bermingham E (2005). A genetic linkage map of the mimetic butterfly *Heliconius melpomene*. *Genetics* **171**: 557–570.

Jiggins CD, Naisbit RE, Coe RL, Mallet J (2001). Reproductive isolation caused by colour pattern mimicry. *Nature* **411**: 302–305.

Joron M, Jiggins CD, Papanicolaou A, McMillan WO (2006a). Heliconius wing patterns: an evo-devo model for understanding phenotypic diversity. *Heredity* **97**: 157–167.

Joron M, Papa R, Beltran M, Chamberlain N, Mavarez J, Baxter S *et al.* (2006b). A conserved supergene locus controls colour pattern diversity in *Heliconius* butterflies. *PLoS Biol* **4**: e303.

Kapan DD (2001). Three-butterfly system provides a field test of mullerian mimicry. *Nature* **409**: 338–340.

Kapan DD, Flanagan NS, Tobler A, Papa R, Reed RD, Gonzalez JA *et al.* (2006). Localization of Müllerian mimicry genes on a dense linkage map of *Heliconius erato*. *Genetics* **173**: 735–757.

Kotiaho JS, Kaitala V, Kolmonen A, Paivinen J (2005). Predicting the risk of extinction from shared ecological characteristics. *Proc Natl Acad Sci USA* **102**: 1963–1967.

Kronforst MR, Young LG, Kapan DD, McNeely C, O'Neill RJ, Gilbert LE (2006). Linkage of butterfly mate preference and wing color preference cue at the genomic location of wingless. *Proc Natl Acad Sci USA* **103**: 6575–6580.

Langham GM (2004). Specialized avian predators repeatedly attack novel color morphs of *Heliconius* butterflies. *Evolution* **58**: 2783–2787.

Lewis DL, Brunetti CR (2006). Ectopic transgene expression in butterfly imaginal wing discs using vaccinia virus. *Biotechniques* **40**: 48–52.

Lewis DL, DeCamillis MA, Brunetti CR, Halder G, Kassner VA, Selegue JE *et al.* (1999). Ectopic gene expression and homeotic transformations in arthropods using recombinant Sindbis viruses. *Curr Biol* **9**: 1279–1287.

Li W, Schuler MA, Berenbaum MR (2003). Diversification of furanocoumarin-metabolizing cytochrome P450 monooxy-

genases in two papilionids: specificity and substrate encounter rate. *Proc Natl Acad Sci USA* **100**: 14593–14598.
Lyytinen A, Brakefield PM, Lindstrom L, Mappes J (2004). Does predation maintain eyespot plasticity in *Bicyclus anynana*? *Proc R Soc Lond Ser B-Biol Sci* **271**: 279–283.
Mallet J, Barton NH (1989). Strong natural selection in a warning-color hybrid zone. *Evolution* **43**: 421–431.
Marcus JM (2005). Jumping genes and AFLP maps: transforming lepidopteran color pattern genetics. *Evol Dev* **7**: 108–114.
Marcus JM, Ramos DM, Monteiro A (2004). Germline transformation of the butterfly *Bicyclus anynana*. *Proc R Soc Lond Ser B-Biol Sci* **271**: S263–S265.
McMillan WO, Jiggins CD, Mallet J (1997). What initiates speciation in passion-vine butterflies? *Proc Natl Acad Sci USA* **94**: 8628–8633.
McMillan WO, Monteiro A, Kapan DD (2002). Development and evolution on the wing. *Trends Ecol Evol* **17**: 125–133.
Mita K, Kasahara M, Sasaki S, Nagayasu Y, Yamada T, Kanamori H *et al*. (2004). The genome sequence of silkworm, *Bombyx mori*. *DNA Res* **11**: 27–35.
Mita K, Morimyo M, Okano K, Koike Y, Nohata J, Kawasaki H *et al*. (2003). The construction of an EST database for *Bombyx mori* and its application. *Proc Natl Acad Sci USA* **100**: 14121–14126.
Monteiro A, Brakefield PM, French V (1997). Butterfly eyespots: the genetics and development of the color rings. *Evolution* **51**: 1207–1216.
Monteiro AF, Brakefield PM, French V (1994). The evolutionary genetics and developmental basis of wing pattern variation in the butterfly *Bicyclus anynana*. *Evolution* **48**: 1147–1157.
Mulder C, Aldenberg T, de Zwart D, van Wijnen HJ, Breure AM (2005). Evaluating the impact of pollution on plant–Lepidoptera relationships. *Environmetrics* **16**: 357–373.
Murray AW (2000). Whither genomics? *Genome Biol* **1**: COMMENT003.
Nijhout HF (1991). *The Development and Evolution of Butterfly Wing Patterns*. Smithsonian Inst. Press: Washington.
Nijhout HF (2003). Polymorphic mimicry in *Papilio dardanus*: mosaic dominance, big effects, and origins. *Evol Dev* **5**: 579–592.
Papanicolaou A, Joron M, Mcmillan WO, Blaxter ML, Jiggins CD (2005). Genomic tools and cDNA derived markers for butterflies. *Mol Ecol* **14**: 2883–2897.
Parkinson J, Mitreva M, Whitton C, Thomson M, Daub J, Martin J *et al*. (2004). A transcriptomic analysis of the phylum Nematoda. *Nat Genet* **36**: 1259–1267.
Reed RD, Gilbert LE (2004). Wing venation and distal-less expression in *Heliconius* butterfly wing pattern development. *Dev Genes Evol* **214**: 628–634.
Reed RD, McMillan WO, Nagy LM (2007). Gene regulation underlying adaptive variation: cinnabar and vermilion in the development and polymorphism of Heliconius butterfly wing patterns (in preparation).
Reed RD, Serfas MS (2004). Butterfly wing pattern evolution is associated with changes in a notch/distal-less temporal pattern formation process. *Curr Biol* **14**: 1159–1166.
Richardson MK, Brakefield PM (2003). Developmental biology – hotspots for evolution. *Nature* **424**: 894–895.
Robertson KA, Monteiro A (2005). Female *Bicyclus anynana* butterflies choose males on the basis of their dorsal UV-reflective eyespot pupils. *P Roy Soc B-Biol Sci* **272**: 1541–1546.
Sheppard PM, Turner JRG, Brown KS, Benson WW, Singer MC (1985). Genetics and the evolution of Müllerian mimicry in *Heliconius* butterflies. *Phil Trans Roy Soc B* **308**: 433–613.
Stavenga DG (2002). Reflections on colourful ommatidia of butterfly eyes. *J Exp Biol* **205**: 1077–1085.
Tobler A, Kapan D, Flanagan NS, Gonzalez C, Peterson E, Jiggins CD *et al*. (2005). First-generation linkage map of the warningly colored butterfly *Heliconius erato*. *Heredity* **94**: 408–417.
Vane-Wright D (2004). Entomology – butterflies at that awkward age. *Nature* **428**: 477–480.
van't Hof AE, Saccheri IJ, Marec F, Brakefield PM, Zwaan B (2007). A high density AFLP-based genetic linkage map for the butterfly *Bicyclus anynana*, covering all 28 karyotyped chromosomes (in preparation).
Wassenaar LI, Hobson KA (1998). Natal origins of migratory monarch butterflies at wintering colonies in Mexico: New isotopic evidence. *Proc Natl Acad Sci USA* **95**: 15436–15439.
Weatherbee SD, Nijhout HF, Grunert LW, Halder G, Galant R, Selegue J *et al*. (1999). *Ultrabithorax* function in butterfly wings and the evolution of insect wing patterns. *Curr Biol* **9**: 109–115.
Xia QY, Zhou ZY, Lu C, Cheng DJ, Dai FY, Li B *et al*. (2004). A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science* **306**: 1937–1940.
Yamamoto K, Narukawa J, Kadono-Okuda K, Nohata J, Sasanuma M, Suetsugu Y *et al*. (2006). Construction of a single nucleotide polymorphism linkage map for the silkworm, *Bombyx mori*, based on BAC end-sequences. *Genetics* **173**: 151–161.
Yasukochi Y, Ashakumary LA, Baba K, Yoshido A, Sahara K (2006). A second generation integrated map of the silkworm reveals synteny and conserved gene order between lepidopteran insects. *Genetics* **173**: 1319–1328.

Chapter 2 - Next generation transcriptomes for next generation genomes using est2assembly.

This Chapter is about the first complete framework for analyzing transcriptomic data to create reference transcriptomes. Aimed at both large sequencing facilities (e.g. University of Edinburgh's Gene Pool service) and small genomic groups, it is responsible for reducing the bioinformatic bottleneck in reference transcriptome generation and standardizing the process. The est2assembly software is used by the InsectaCentral database presented in later chapters.

Citation

Papanicolaou, A. et al., 2009. Next generation transcriptomes for next generation genomes using est2assembly. BMC bioinformatics, 10(1), 447.

BMC Bioinformatics

BioMed Central

Software

Open Access

Next generation transcriptomes for next generation genomes using *est2assembly*

Alexie Papanicolaou*[1,2], Remo Stierli[3], Richard H ffrench-Constant[2] and David G Heckel[1]

Address: [1]Department of Entomology, Max Planck Institute for Chemical Ecology, Jena, Germany, [2]School of Biological Sciences, Centre for Ecology and Conservation, University of Exeter, Penryn, UK and [3]Department of Computer Science and Statistics, University of Rhode Island, Kingston, USA

Email: Alexie Papanicolaou* - alexie@butterflybase.org; Remo Stierli - rstierli@cs.uri.edu; Richard H ffrench-Constant - r.ffrench-Constant@exeter.ac.uk; David G Heckel - heckel@ice.mpg.de

* Corresponding author

Published: 24 December 2009

BMC Bioinformatics 2009, **10**:447 doi:10.1186/1471-2105-10-447

Received: 3 August 2009
Accepted: 24 December 2009

This article is available from: http://www.biomedcentral.com/1471-2105/10/447

Abstract

Background: The decreasing costs of capillary-based Sanger sequencing and next generation technologies, such as 454 pyrosequencing, have prompted an explosion of transcriptome projects in non-model species, where even shallow sequencing of transcriptomes can now be used to examine a range of research questions. This rapid growth in data has outstripped the ability of researchers working on non-model species to analyze and mine transcriptome data efficiently.

Results: Here we present a semi-automated platform '*est2assembly*' that processes raw sequence data from Sanger or 454 sequencing into a hybrid *de-novo* assembly, annotates it and produces GMOD compatible output, including a SeqFeature database suitable for GBrowse. Users are able to parameterize assembler variables, judge assembly quality and determine the optimal assembly for their specific needs. We used *est2assembly* to process *Drosophila* and *Bicyclus* public Sanger EST data and then compared them to published 454 data as well as eight new insect transcriptome collections.

Conclusions: Analysis of such a wide variety of data allows us to understand how these new technologies can assist EST project design. We determine that assembler parameterization is as essential as standardized methods to judge the output of ESTs projects. Further, even shallow sequencing using 454 produces sufficient data to be of wide use to the community. *est2assembly* is an important tool to assist manual curation for gene models, an important resource in their own right but especially for species which are due to acquire a genome project using Next Generation Sequencing.

Background

Much of the recent progress in our understanding of genomics has come from the study of model genetic organisms such as the fruit fly *Drosophila melanogaster*, the nematode worm *Caenorhabditis elegans* and the zebrafish *Danio rerio* [1]. In these model species, a full genome sequence combined with a well annotated collection of gene models is currently available. To address fundamental questions in evolutionary biology, however, we need to expand the genomic and transcriptomic resources avail-

BMC Bioinformatics 2009, **10**:447 http://www.biomedcentral.com/1471-2105/10/447

able for non-model species, notably insects [1]. Non-model insects represent attractive models for the study of a range of important biological questions both of applied and fundamental importance, such as those relating to studying insecticide resistance [2], wing pattern development [3] or co-evolution [4]. From a functional biologist's point of view, crucial experimental tractability can be gained via a combination of rich sequence data (including Expressed Sequence Tag - EST - collections from several tissues), gene models, functional annotation and in-depth knowledge of an organism's genetics, preferably coupled with the ability to manipulate them [5,6].

Since the advent of EST sequencing, the number of organisms represented in dbEST (which houses all public Sanger-sequenced EST projects) [7] has exploded. Transcriptomics has grown to be at least as an important resource to non-model species communities as it has proven for the traditional models. It is now conceivable that many non-model species will have at least a workable outline of their genomes available. In turn, large collections of ESTs important in genome annotation will be generated as the community identifies -omics as a major resource for species with an evolutionary importance [8-10]. This scenario is already a reality as the related technologies become more cost-effective. Such non-model species transcriptome projects, being focused on particular applications and biological questions, are especially useful to the wider community even if not targeting annotation of any specific genome [11-13]. One of the most important benefits of shallow EST sequencing is the ability to acquire candidate sequence data for downstream applications such as phylogenetics, multi-locus population genetics and expression studies [14]. It is hoped that with a wide application of Next Generation Sequencing (NGS) the bottleneck in obtaining sequence data will no longer exist [15]. The reality is, however, that this vast amount of sequence data has outstripped the ability of most researchers working on non-model species, who often have limited bioinformatic support, to analyze and mine their new datasets. Further, there is currently no standardized platform for providing researchers with such analyses for transcriptomes. The Generic Model Organism Database (GMOD) group, derived mainly from the model species communities, is a collection of software, platforms and standardized approaches in dealing with -omic data. They are best known for the GBrowse project, the sequence viewer used by WormBase, FlyBase and others [16]. One of the most important contributions of the GMOD group, however, is the development and dissemination of standards capable of generic use: extensive use of BioPerl [17], database connectivity frameworks and Chado, the generic database schema for almost any type of data produced by the community [18]. Such standards allow tools to be capable of a unique level of interconnectivity and interoperability.

Here we describe a software suite, *est2assembly*, which aims to address the above bioinformatic deficit while being embedded in the GMOD framework. It accepts raw sequence data (from 454 or Sanger technologies) and produces annotated assemblies in a GMOD/Chado-compatible format with minimum user input. Further, using the common file format of GFF (which stands for General Feature Format), users can share their data with collaborators and visualize them with tools such as GBrowse. The platform is highly automated and standardized, and we show it allows for direct comparison between various datasets. The modular nature of *est2assembly* allows users to independently make use of different subroutines. Extensive log files guide the user through the assembly process and the output. We demonstrate this platform using a range of 454 data from a phylogenetically diverse sample of insects. We benchmark the platform and compare these non-model species collections with 745,124 public EST data from *D. melanogaster* collected via conventional capillary sequencing which is still considered the gold standard in insect EST data and *Bicyclus anynana*, the butterfly with the highest number of Sanger-sequenced ESTs in GenBank.

Implementation

Software dependencies

All software on which the platform depends is free and installation requirements are straightforward. In brief, the platform makes extensive use of BioPerl (1.6+) and other open-source Perl modules available from CPAN. The GPL-licensed MIRA assembler is required [19]. The proprietary Newbler2 - including the associated SFF toolkit - (454 Life Sciences) is optional. For 454 next generation sequencing files we use sff_extract [20] as this is the only known 454 basecalling software which is not under a restrictive license. In this paper, we made use of the complete platform and utilized both Newbler2 and MIRA version 2.9.37. Further, some modules have dependencies such as installation of the Chado database schema [18], EMBOSS [21], NCBI-BLAST [22], SSAHA2 [23], RepeatMasker [24] (with an recommended registration to RepBase [25]), prot4EST [26], FASTY 3.4 [27] and annot8r [28]. Due to the modular nature of the platform, a researcher needs to install only the components which will be of use their application of the platform. We provide a comprehensive installation script to ease the procedure of installation of the above 3rd party software.

Read pre-processing

The est_process module is driven by *preprocessest.pl* which accomplishes the following steps: i) project creation, including calling the bases or reading the SSF files; ii)

masking short sequences (e.g. adaptors) using SSAHA2; iii) BLAST2 to detect unwanted sequences (as defined by the user) which can cause problems in the assembly (e.g. mitochondrial, rDNA and contaminants); iv) removal or masking using these BLAST output files; v) Repeat masking performed using RepeatMasker and a user provided database (which can be extracted from RepBase and customized); vi) polyA/T screening performed using a tiered approach of a custom algorithm. Then a final step cleans the output files and prepares assembly specific input files including an XML NCBI TraceInfo file. A user may interrupt the program and resume from any of the above steps.

The conversion of raw trace files to input files for an assembly is performed independently for each technology. Users have the option to convert and quality trim Sanger-derived sequences using the gold standard of phred or in the case of data derived from ABI sequencers data, make use of any internal KB basecalling. In the latter case, we use a custom quality trim subroutine provided by Steffi Gebauer-Jung (MPI for Chemical Ecology) involving two sliding windows to avoid local optima: a larger one scans the trace for a sudden drop in quality values and a finer search pinpoints the exact location. For 454 sequencing, due to licensing prohibiting the use of Roche's proprietary flowgram extracting software by ordinary researchers, SFF files are extracted using sff_extract, an open-source alternative. In addition to detection of low quality regions, we identify adaptor sequence introduced either in the making of the cDNA library or subsequent pre-sequencing steps. We use a two tiered search using SSAHA2, which combines the SSAHA and the cross_match algorithms [29] and BLAST2 from NCBI. We found that the SSAHA2 search itself is best utilized by using three iterations: two searches for adaptor sequence (with one prior- and one post-polyA/T masking) and one restriction site search with a parameterization for extremely short target sequences (restriction sites tend to be ca 7-10 bp). The platform uses BLAST2 to screen for common contaminants found in molecular biology labs via a customizable FASTA database. In any transcriptome project, it is also undesirable to clutter the assembler with sequences that are overrepresented due to transposon activity, mitochondrial or rDNA origin. The platform allows users to screen them using RepeatMasker via a user-defined database. To assist Lepidopterists in particular, we have included a prediction of repetitive elements from *Bombyx mori* (C. Smith, University of San Francisco, pers. communication) which users can concatenate with the Insect repeat library from RepBase. The intensive steps of running BLAST and SSAHA2 can seamlessly utilize multiple threads to reduce run time.

Trimming of polyA/T tails is essential in EST projects and we use a routine to iteratively scan for such homo-oligomers. The routine can also be used independently of the platform and is highly customizable; users can specify which base (A, T, C or G) they wish to search for, seed length, min/max length of homo-oligomer, depth of search of each sequence and other options. In order to minimize false positives in A/T rich genomes or errors produced by the pyrosequencing methodology of 454, the platform utilizes 5 rounds with increasing minimum length and decreasing search seed length. Except for the second round pair, the scan is performed only at the ends of the sequence for a length of one-third of the total sequence length. The second round scans deeper to 350 bp from each end. An additional feature of the routine is the use of any suspected polyadenylation signal site (PAS) upstream of a hypothesized polyA/T site. Currently, we use a simple pattern search but implementing a model-based approach [30] could be of use. This assists in correcting the masking and avoiding over-trimming of the 3' UTR or for allowing a short polyA sequence which would normally be below acceptable length to be masked. In addition, it uses this information to detect false positives if there is a significant amount of sequence (>50 bp) between the end of the polyA and the start of any vector masked sequence (or the end of the sequence). If no PAS site is found upstream of the polyA tail (in a 50 bp window) and the suspected polyA is shorter than the specified cut-off, then the polyA is not masked but still tagged in the log-files. We found that this option enhances polyA masking in Sanger-derived sequences but by default is switched off for short reads. An output for each polyA/T found and which criteria were used is produced in a log file should users wish to exploit them in gene model construction. At this stage, an XML file (using the NCBI TraceArchive template) containing the low quality, adaptor sequence and polyA/T trim points is generated. This file, when used with the original untrimmed and unmasked file, can guide assemblers such as MIRA on how to perform clipping of undesired regions. This approach can be used to tackle any potential false positives that may arise in the preprocessing steps.

It is worth noting here that we find that near-perfect signatures of the 454 adaptor sequences can persist even within regions of high quality assembly, which could be the result of the chimeric ligation of molecules. In such a scenario, we recommend that if manual inspection is not feasible that the sequence region is masked and the assembler allowed to determine if the region is truly an adaptor sequence or part of the sequenced species genome: false positives will have multiple reads in the assembly exhibiting high identity down- and up-stream of the suspected site and therefore still assemble. To facilitate assemblers who cannot make such judgements, preprocessest.pl allows the flagging of sequences which have more adaptor sequences than the user defines. If the user

BMC Bioinformatics 2009, **10**:447 http://www.biomedcentral.com/1471-2105/10/447

wishes then only the longest stretch of high quality sequence between two adaptors is used.

Optimal assembly

The output files of est_process are provided to the second module, parameterize_assembly. It could be straightforward to plug-in various assemblers in future versions but we currently make use of Newbler as it is the standard for 454 data and MIRA due to its ability to analyze Sanger/Next Generation data concurrently and the provision of excellent support. In the current version, datasets from Sanger and/or 454 can be provided and users will concatenate any datasets originating from identical technologies. A configuration file is responsible for defining which parameters are passed on to the assembler MIRA. This allows for multiple runs of the same datasets in order to explore the parameter space.

Assembly quality is estimated using analyze_blast.pl via summary indexes based on the coverage of one or more reference organism databases used in a similarity search (e.g. NCBI-BLAST). Coverage is calculated on a base pair basis by counting unique hits to a particular base pair. Overlapping coverage (i.e. redundancy) is calculated by counting the total number of hits a base pair (or amino acid; the platform uses the term position) receives. We perform these calculations for both the assembly and the database. The ratio of redundancy over coverage is summed for the database and the assembly. If more than one reference database is used (which is recommended in order to discount any organism-specific effects) then the total sum is used. One has to be aware that the absolute numbers of coverage are volatile as they are dependent on the BLAST cut-offs used. When the same cutoffs are used, however, comparison of assemblies is a meaningful index to evaluate how parameterization influences the assembly. Another quality control index is the number of reads included in the assembly. This can act as a proxy for the downstream utility of an assembly, for example the number of SNPs which can be determined or the likelihood we can detect alternative splicing or frame-shift-causing sequencing errors. Finally, we also consider the proportion of the reference database covered (or the average if more than one is used) as a proxy to eventual gene finding.

In EST assemblies a large portion of the consensus can be non-coding sequence. Such sequences often diverge rapidly due to the lack of any selective constraints. The unfortunate result in any assembly process is that sequences of this type, which are from identical genomic regions, fail to assemble together. The script trim_assembly.pl is one of the two methods to remove such redundant contigs. It first scans for polyA/T tails which may have been built during the assembly. Then it defines a set of 'high-quality' contigs using user-specified cut-off for length and number of reads included. The other contigs are only included in the final set if they a) don't have a high sequence similarity to a high-quality contig, b) have a high sequence similarity to a reference proteome, c) specifically requested by the user by proving a list file. The output of the contigs which have been excluded is cataloged in a log file.

Protein identification, SNP discovery and data mining

Data mining of EST projects is driven by searching the dataset for the signature of a favorite protein or sequence. The platform makes use of BLAST similarity searches using the contig consensus. The analyse_assembly.pl and analyze_blast.pl scripts, which are employed during parameterization, can be used standalone to estimate the quality indexes for any BLAST report. They allow users to identify the exact coverage of a FASTA input file in relation to a reference database using BLASTx, BLASTn or tBLASTx. They provide the ability to run multiple blasts in a threaded fashion with one command. The script has the additional ability to output a FASTA file with the part of the input file which matches the reference and a second file with the part which did not. This approach is very useful in extracting the segment of the assembly which is known to be coding. For example, a tBLASTx approach is useful when multiple species have been sequenced but no reference proteome exists or is under-annotated (see Results section).

Deeper annotation can be performed using predicted proteins. Protein predictions are accomplished using prot4EST [26] (included in the distribution). The current version of prot4EST does not produce the Open Reading Frame (ORF) which can differ substantially from the contig (one of the utilities of prot4EST is that it corrects for frameshifts). We acquire the ORF using FASTY or failing that, EMBOSS's transambig and attempt to correct for any ambiguous codons to match the consensus. These are then annotated using similarity to known proteins which may have annotated ontology terms: Gene Ontology (GO) [31], Enzyme Commission (EC) [32] and Kyoto Encyclopaedia of Genes and Genomes (KEGG) [33]. Annotations based on electronic similarity are assigned using annot8r. If computational resources allow, InterProScan annotations can be included to allow users to search for specific protein domains, motifs and sequence signals. The output of the above procedures is then converted to a common file format using ic_annot8r2gff.pl and can be databased.

Further, we provide the script *analyse_assembly.pl* to help with BLAST reports on a single computer but if EST projects are commonplace then access to a PC-farm or a high performance computing system is required. For this reason, we include a set of simple scripts to facilitate use

BMC Bioinformatics 2009, **10**:447 http://www.biomedcentral.com/1471-2105/10/447

and error-checking of LSF queue submissions allowing, therefore, for the automation of BLAST and InterProScan annotations.

Of particular interest to biologists is the identification of Single Nucleotide Polymorphism markers (SNPs). The assembler MIRA produces such a set and can be included in the assembly GFF file. We also extract a 'high-quality' SNP dataset by including those SNPs which have the minor allele frequency above a user-specified threshold and have at least 20 invariable bases up and downstream of the SNP position. This padding can be customized by the user but we default to 20 in order to design primers and create a unique identification sequence for submission to dbSNP [34]. This SNP identification is accomplished via ic_create_snps.pl which estimates the position of each SNP in relation to the ORF and, if determined as coding, provide the codon position, the alleles and whether the nucleotide change causes a synonymous or non-synonymous amino acid change. Due to prot4EST's approach to determine the ORF, a simple translation of co-ordinates is not possible. We, therefore, perform a local alignment using FASTY [27]. We also include a similar implementation for SEAN [35] which would be of use to users with small amounts of data.

We can utilize GFF as the middleman to populate Chado and Bio::DB::SeqFeature (SeqFeature) database schemas. In this paper, we focus on the simple format and speed advantages of SeqFeature as most users will be dealing with a limited number of datasets. We provide a set of scripts that produce both Chado- and SeqFeature-compatible GFF files for each data type which the platform is capable of processing: CAF assemblies in read-contig or contig-read sorting order, BLAST reports, ORF predictions, SNPs predictions and KEGG, GO and EC annotations from annot8r. As the proper linking of the various reference sequences and their features is essential, we have a specific strategy for creating the GFF files for use with GBrowse (Figure 1C): a contig view is composed of a reference contig and associated annotation features such as the assembled EST data, SNP markers, BLAST annotation etc. is anchored to it. The ORF feature is also anchored to the contig but also exists as a separate reference sequence in a second web page. This ORF object has its own associated annotation, including a polypeptide features which serves as an anchor for the protein prediction reference. Likewise, this protein prediction allows for anchoring protein-based annotations such as estimated molecular weight, assigned ontology terms and other protein-based data in a third web page. This approach enforces the one-to-one relationship between an ORF and a protein but allows for one-to-many relationships between an assembled contig and protein predictions.

Benchmark datasets

Data for benchmarking the platform was provided by dbEST (for *D. melanogaster* and *B. anynana*) or by collaborators. In more detail, the GS20 sequencing of a *Manduca sexta* (tobacco hornworm moth) hemocyte cDNA library was provided by Haobo Jiang and is published by Zou *et al* [36]. The *Chrysomela tremulae* (a beetle) midgut GSLFX and *Manduca sexta* midgut are published by Pauchet *et al* [37] and [38] respectively; the GSFLX of cDNA from whole larvae of *Euphydryas aurinia* (marsh fritillary butterfly) was provided by Yannick Pauchet (University of Exeter); the GSFLX-Titanium sequencing of cDNA from wing discs of *Papilio dardanus* (African swallowtail butterfly) by Iva Fuková (University of Exeter); the Sanger capillary and GSFLX data of *Heliconius melpomene* were prepared from wing-discs of developmental stages from late larval through to mid-pupal stage by Ronald Lee and Chris D. Jiggins (University of Cambridge) and is published by Ferguson *et al* [39]. The Sanger capillary and GSFLX data of *Heliconius erato* was also generated from wing-discs and was provided by W. Owen McMillan (North Carolina State University). The GS20 *Melitaea cinxia* dataset from whole larvae is published by Vera *et al* [40] and was downloaded from NCBI's Short Read Archive after communication with Howard Fescemyer (Pennsylvania State University). The *Bicyclus anynana* data are based on the capillary-sequencing technology of a variety of tissues, were obtained from dbEST and is published by Beldade *et al* [41]. The *D. melanogaster* data are also based on capillary-sequencing technology of a variety of tissues and were retrieved from dbEST. We did not include any singletons in the resulting assemblies. For the saturation curves, we used the *H. melpomene* 454 preprocessed dataset (without any Sanger sequences) and created pseudo-datasets by randomly splitting it in datasets containing 1/5, 2/5, 3/5 and 4/5 of the initial data. We repeated the procedure 5 times for each pseudo-dataset thus generating and annotating 20 pseudo-assemblies using the same procedures as for the main assembly.

Reference proteomes used in this study were from *D. melanogaster*, *Anopheles gambiae*, *Apis mellifera*, *B. mori* and *Tribolium castaneum*. For each species we used the RefSeq [42] and UniProt [43] curated proteins, concatenated with the predictions provided by each organism's Genome Database [44-50] and made non-redundant at the 100% level using cd-hit [51]. In the cross-dataset comparison, we identify ORF coverage by calculating the proportion of the reference proteome aligned to the assembly (e-value <= 1e-5; bit-score >= 80 bits) as an indication to gene-finding. We also estimate the proportion of the assembly aligning to the proteome (e-value <= 1e-5; bit-score >= 80 bits) to determine the portion of the assembly likely to be coding and also include the improvement of including a tBLASTx search (e-value <= 1e-15; bit-score >= 80 bits).

BMC Bioinformatics 2009, **10**:447 http://www.biomedcentral.com/1471-2105/10/447

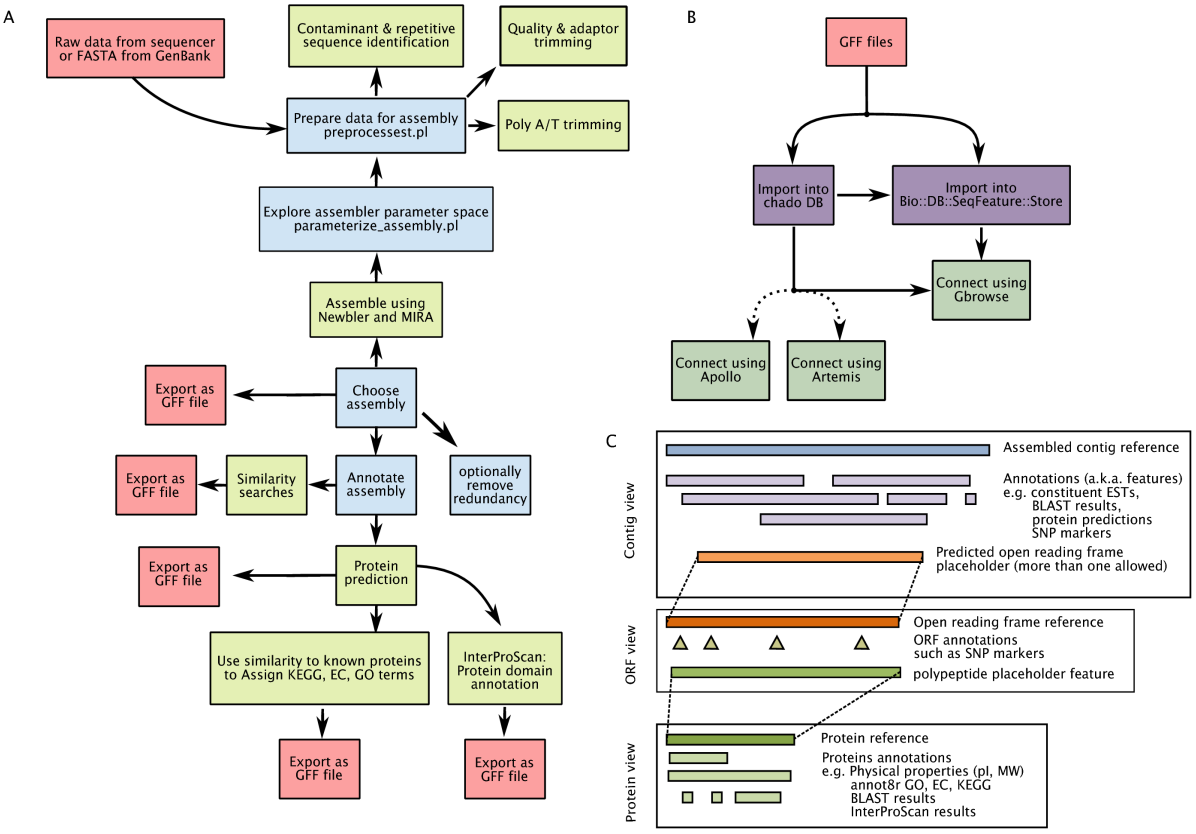


Figure 1

Schematic diagram of the *est2assembly* platform. (A) Sub-routines processing and annotating the EST data. Note that all outputs are in the common GFF standard and therefore can be accessed by GMOD-compatible software. (B) Diagram illustrating the the ability of *est2assembly* to produce a GBrowse sequence view. (C) Diagram illustrating a triple page approach to graphical outputs from *est2assembly*: First, a page showing the assembled contig and associated annotation, second, a page showing each predicted ORF and its annotation and, third, a page focused around the annotated protein object. Note that each page is linked and allows for rapid navigation to genes of interest.

Implementation overview for the biologist

The *est2assembly* platform allows for the processing of data either directly from DNA sequencers (pyrosequencing or Sanger based) or as FASTA files. The software also allows users to combine their own datasets with those available in public databases and a script is included to automatically download such sequences from the European Bioinformatics Institute (EBI). Moreover, any large sequenced genomic regions (such as from Bacterial Artificial Chromosomes; BACs) can contribute CoDing Sequences (CDSs) to the assembler. This form of dataset concatenation has the advantage that a pool of shorter NGS reads will assemble better if a longer sequence (such a full-length mRNA sequence) is also included in the same pool. Currently, two sequencing technologies can be processed: Sanger capillary-based data and 454 pyrosequencing data. The input data is fed into the preprocessest.pl which removes low quality sequences, any adaptors and polyA/Ts that may be present. This script also removes any contaminants and repeats which can cause serious misalignments. Two versions of the processed FASTA files are produced: trimmed and 'masked', the latter accompanied by a quality file and an NCBI Traceinfo XML file which defines trim points in relation to the original files. Researchers can choose to use the untrimmed but masked files or the original sequences (with the XML file which contains the exact regions of high quality sequence) or the trimmed files for assemblers such as Newbler. Further, at this point in the pipeline, graphs can also be generated (using two scripts provided in the distribution) to examine the effect of the pre-processing on the data.

Once a user is satisfied with the quality control of the reads, the assembly can begin. An optimal assembly needs

BMC Bioinformatics 2009, **10**:447 http://www.biomedcentral.com/1471-2105/10/447

to be chosen according to the needs of a project but often the assembler is used as a black box despite the fact that assemblers are mere computation machines and therefore the results may or may not be biologically meaningful. For this reason, the next step is submitting these files to a second script, parameterize_assembly, which launches the assemblers (currently Newbler and MIRA are supported) with varying parameters, compares the results to one or more reference proteomes and computes a number of indexes suitable for transcriptome projects. Which index is most useful (i.e. the optimality criteria) depends on the aim of the particular project. For example, in a gene-hunting project one may wish to optimize for the number of genes discovered and minimize redundant contigs, where as in SNP project, one may wish to maximize for the number of reads in the assembly and tolerate redundancy which can later be addressed manually for contigs of interest. For the reference proteomes, we suggest that more than one is provided in order to remove species-specific bias and increase the power of detecting coding sequences but, for computational reasons, too divergent species will not be useful for parameterization. In our work and this paper, we used species with a genome sequence which are in the same Phylum as our data datasets. As parameterization is focused on exploring how different parameters behave, the exact details of the reference proteome and BLAST cut-off values are not as important since the platform ensures they are used in a consistent fashion.

Simple BLAST-driven indexes are reported for each proteome: the number of queries which have similarity with a reference protein; the number of reference proteins which have similarity to a contig in the assembly. It then calculates the indexes for each base pair/amino acid rather in order to detect the level of partial ORF sequencing. Further, an annotation redundancy index is based on the number of one-to-many hits (summed in both directions) between the reference proteomes and the assembly. Finally, *est2assembly* reports the number of reads included in the assembly. Which criterion is chosen to identify the best assembly is left to the individual researchers. In this paper, since we are focused on maximizing the 'number of genes found' - like many non-model species projects - we used the reference proteins found as the main criterion.

Ignoring the parameterization step is not recommended, as Figure 2 shows: the MIRA.a0 parameter set is the default for MIRA's 'accurate quality' setting but produces a suboptimal assembly using the criteria of annotation redundancy. One should note that even though transcriptome sequencing (especially with NGS technologies) is producing transcripts from the whole mRNA, it is unlikely that full length transcripts are sequenced. Figure 3 shows that proportion of identified proteins in terms of CDS coverage was significantly lower than the proportion in terms of number of identified proteins (e.g. for MIRA 20% vs 42%). In addition, for this dataset, MIRA outperformed Newbler with our chosen criterion, showing it is important to attempt an assembly with more than one assembler.

Once the desired assembly is chosen, users may opt for the removal of a subset of redundant contigs using trim_assembly.pl and evaluate if there is a loss of 'number

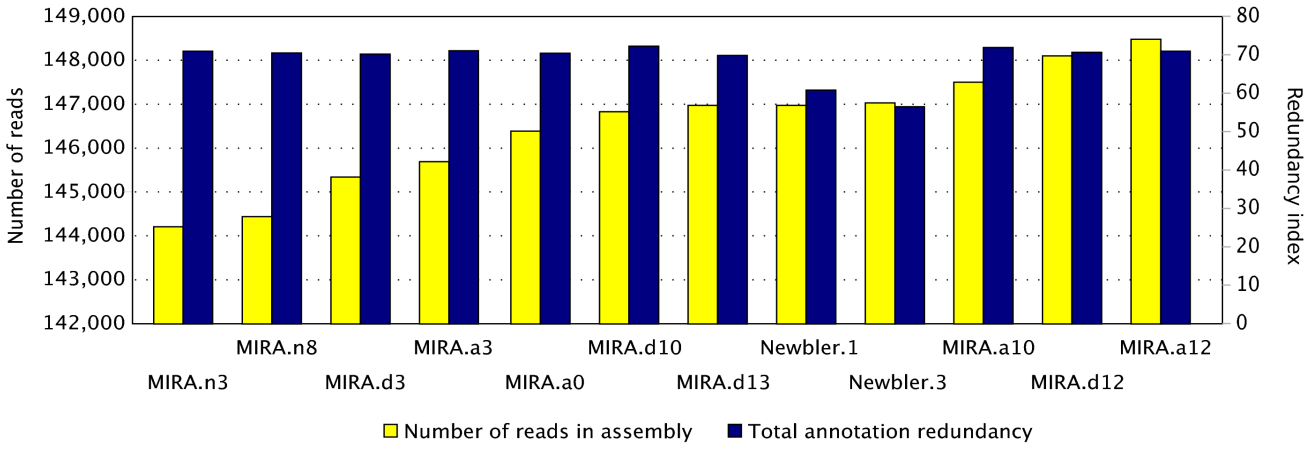


Figure 2
Exploration of the parameter space on the *E. aurinia* dataset. Effect of parameterization on assembly is significant. In this dataset, Mira.a0 is the default settings for an 'accurate' assembly. One benchmark is number of reads as lower number of reads result in lower coverage. Another is the redundancy index estimates the level of one-to-many edges (in both directions) exist in an alignment graph between an assembly and the same reference proteome. Newbler seems to outperform MIRA if annotation redundancy is the estimator but see Figure 4.

BMC Bioinformatics 2009, **10**:447 http://www.biomedcentral.com/1471-2105/10/447

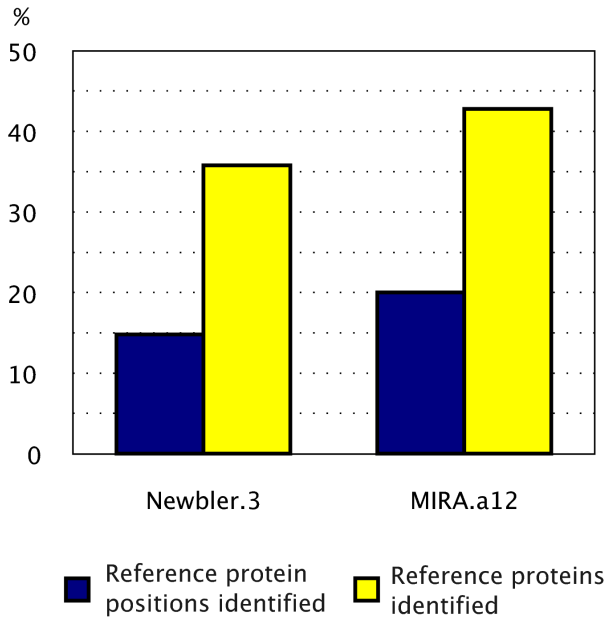


Figure 3
Comparison of the Newbler.3 and MIRA.a12 assemblers with respect to the numbers of amino acid residues or proteins identified via the *est2assembly* pipeline. In this *E. aurinia* dataset, we used the BLASTx similarity to *Bombyx mori* (cut-off 50 bits) in order to compare performance. MIRA produces an assembly which identifies more of the reference proteome. Further, at this coverage, we do not have a complete coverage of each gene as the proportion of individual amino acids identified is lower (see text for discussion). As this project is a gene-finding one, we choose the MIRA assembly for downstream application.

of genes found'. By redundancy we mean here the fraction of contigs that are likely to originate from the same locus - as defined by the degree of similarity. Often, an assembler fails to align them due to a high number of mismatches caused by a non-conserved region (such as in non-coding regions) accumulating mutations, an alternative exon or - in libraries constructed from outbred individuals- multiple SNPs. The aim of a reduction of redundancy is to reduce the strain of computing resources on the subsequent annotation steps. Researchers will, therefore, annotate a considerably smaller dataset. We support GFF conversion for a number of publicly available tools that we consider to be of most use in transcriptome project and we use routinely: the CAF format (the new standardized file format for assemblies); prot4EST (ORF prediction); BLAST (similarity annotation); annot8r (Gene Ontology, Enzyme Commission and KEGG pathway term assignment according to similarity to known proteins) and InterProScan (protein domain identification).

As our interest is primarily in ensuring that data produced by multiple researchers can be integrated, we decided to utilize the community tools of BioPerl and GMOD. The bioinformatics community has been converging on a set of standard formats. One such flat-file format for sequence annotation is the General Feature Format (GFF) specification which is currently being standardized as version 3. The GFF format is a tab and semicolon delimited file making it both machine- and spreadsheet- readable and has become the format of choice for the GMOD software group. From a database perspective, there are two additional important formats: BioPerl's Bio::DB::SeqFeature and Chado. The former is a highly denormalized database schema which allows for rapid queries of sequence data by sacrificing control of data integrity. Chado, on the other hand, is a normalized modular database schema created to serve as the main data warehouse of multiple types of data. It is logical therefore for researchers to utilize Chado as a data warehouse and Bio::DB::SeqFeature for driving user-visualization software such as GBrowse. Data can be loaded into a database via BioPerl and we provide a script to load multiple GFFs in the correct order and allow for later additions. The Chado schema requires a PostgreSQL database and we find that the SeqFeature database works well with PostgreSQL as well. Once the database is loaded, one can use to drive popular tools such as GBrowse, Apollo and Artemis (Figure 1B) in order to curate the project. Transcriptome project curation requires the ability to join contigs and we find that the user-friendliness of the proprietary program Geneious (Biomatters Ltd, Auckland New Zealand) is efficient for the purpose and a free version is available. A future version of this platform may make use of Geneious' interactivity interface (API). Due to the popularity of GBrowse as a sequence viewer, we provide configuration files that can be readily customized. When the complete analysis is loaded, researchers and their collaborators can view any annotated contig in three inter-linked web pages: assembled contig, predicted ORF and protein (Figure 1C).

Results

As est2assembly is unique in the field as it is not one pipeline but a complete framework to analyze transcriptomes from raw data to a format wet-lab biologists can analyze. The preprocessing step has been built to take advantage of NCBI's TraceXML in order to annotate vector and clipping positions. This allows us to use the original (unmasked) data to produce the assembly using MIRA which can make intelligent decisions if a clipped region is a false positive. The assembler parameterization allows bioinformaticians to seamlessly explore the parameter space as Figures 2 and 3 mentioned above (Implementation overview for the Biologist). Had we used the default settings, we would have an assembly that had a lower number of identified

BMC Bioinformatics 2009, **10**:447 http://www.biomedcentral.com/1471-2105/10/447

reference proteins (data not shown), a lower number of reads used but similar degree of the redundancy index (Figure 2). Likewise, had we opted for the Newbler assembler without investigating the MIRA assembly, we would have both a lower number of proteins identified and shorter CDSs (Figure 3).

In addition, as we mention above and show below, current NGS assemblers produce non-redundant contig sets. This is a correct procedure in order to avoid assembling close paralogs (as joining contigs is easier that splitting them) but results in a downstream computational problem with annotating a redundant set of objects. Our trim_assembly approach is highly customizable using concepts intuitive to biologists and produces better clustering than the standard cd-hit-est we used to use: for one of the more redundant assemblies we started with 54,748 contigs and reduced it to 37,012 contigs with trim_assembly when compared to 51,012 when using cd-hit-est with both strand search enabled.

Subsequently, we extend the usefulness of prot4EST by allowing users to build a ORF model even for species where no ORF is annotated in the public domain. Further, our SNP pipeline uses a similar approach as SEAN but predicts more markers (Table 1) and is built to be fast and efficient with a large number of data (e.g. identification and ORF classification of SNPs in an assembly needs ca. 60 minutes for 14,817 contigs with 246,477 sequences when SEAN needed more than one day due to high I/O usage) but also to decrease the number of SNPs which would be useful to wet-lab biologists: with the 454 technology we have more candidate SNPs that biologists can afford to screen or make use of (see Table 1 for comparison with MIRA which is a liberal predictor). Manual inspection is essential in order to identify markers which are most useful in downstream genotyping methods. For example, any base covered with less than 4 reads is of no use for a SNP call as one cannot distinguish a SNP from a sequencing error. In addition, the platform predicts high quality SNPs by demanding a certain region surrounding the SNP to be invariable in order to assist with primer design. Further, by comparing with the position in the predicted ORF, a marker is classified as non-coding/coding and then determined if causing a synonymous (amino acid is preserved) or non-synonymous (amino acid is changed) mutation. Perhaps, however, the single most important innovation in the field of transcriptome processing is the utilization of the GFF file format and integrating the assembly, protein predictions and annotations into a format the GMOD framework.

Utility

We used a diverse dataset to test and build this platform and due to the standardized approach which it follows, we have been able to evaluate data from different sequencing technologies and protocols in order to offer insights on non-model species transcriptomics. The cost-

Table 1: SNP marker identification in trimmed assemblies

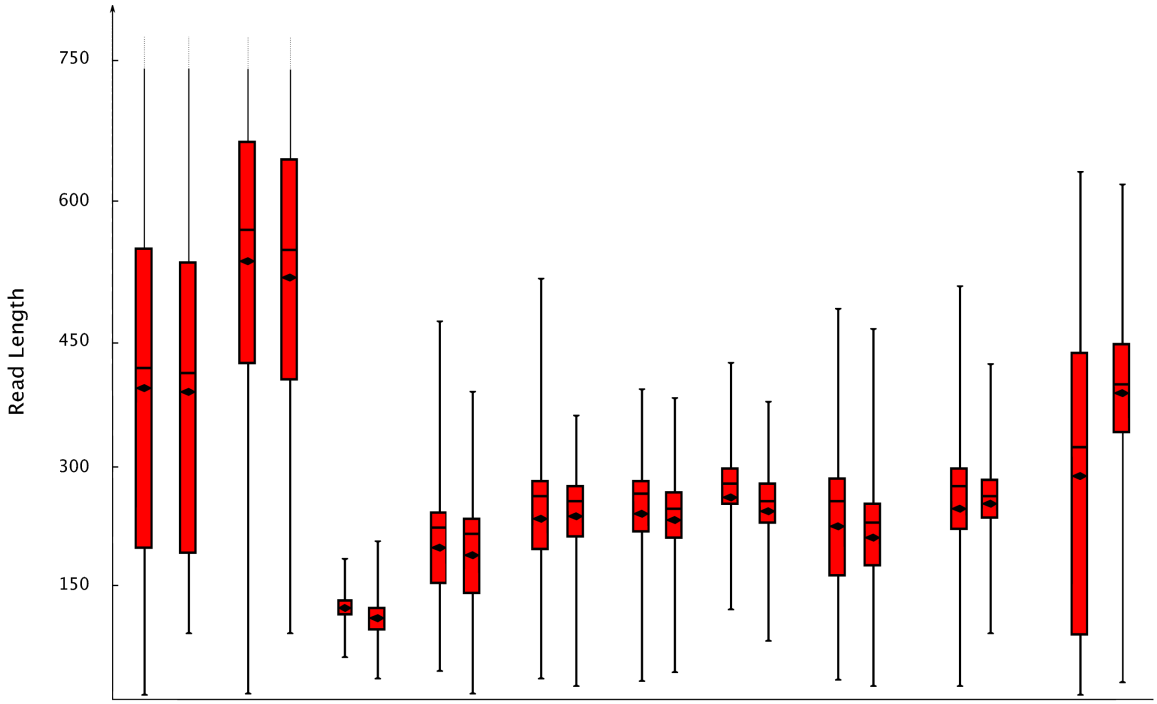
Dataset	Trimmed contigs	Total High Quality SNPs	Coding SNPs	Synonymous SNPs	Predicted with SEAN	Predicted with MIRA
D. melanogaster (Sanger)	24,629	77,969	49,341	15,535	8,060	415,501
B. anynana (Sanger)	11,942	18,271	10,783	4,773	3,282	16,847
M. cinxia (GS20)	12,492	5,622	3,521	1,979	5,918	Not estimated
M. sexta (Sanger + GSFLX)	12,635	7,593	5,026	2,594	6,510	527,469
C. tremulae (GSFLX)	9,771	3,238	2,087	978	2,905	Not estimated
E. aurinia (GSFLX)	8,984	6,132	3,921	2,170	5,543	Not estimated
H. erato (Sanger + GSFLX)	12,130	8,720	4,893	2660	8,744	22939
H. melpomene (Sanger + GSFLX)	16,631	65,047	28,536	18,613	27,526	4,090,305
P. dardanus (GSFLX Titan.)	25,083	49,421	20,694	11,069	7,061	Not estimated

BMC Bioinformatics 2009, **10**:447 http://www.biomedcentral.com/1471-2105/10/447

effectiveness of NGS has been a primary motivation for non-model species researchers to initiate project yet others are worried about the quality of data resulting from such short reads. Examination of the data after preprocessing is essential in order to make a meaningful comparison and our graphical tools allows researchers to compare their raw data with other datasets (Figure 4).

In gene discovery projects, if one hopes to provide an accurate level of annotation, the length of the contigs is an important element which must be considered during project design. As Figure 5 shows, the number of reads increases significantly the length of contigs (for example when comparing *H. melpomene* with all the other GSFLX datasets) but technology has the largest effect. GS20 seems to be of limited use and the newest GSFLX-Titanium has comparable contig length to 2.5 GSFLX runs. Further, with increased read number and length, we get an increased contig coverage. The coverage of each contig (how many sequencing reads are assembled together in one contig) is an important limiting factor in SNP prediction, error correction and the overall quality of the assembly. For population genomicists, substantial contig coverage offers an additional advantage in being able to estimate the frequency of a particular SNP marker. Further, low frequency non-synonymous SNPs can guide a curator to regions of misassembly or erroneous ORF prediction. Once verified to be true non-synonymous SNPs, curators can look for genes showing an excess of non-synonymous polymorphisms and, therefore, possibly evolve under balancing selection. Even though, current users need to perform this latter step manually, it would be of use to automate it for assemblies which are hand-curated. For such an investigation to be profitable, however, the design of the project, especially how many and which individuals, to be included in the cDNA library must be carefully planned.

The number of contigs is, however, rarely an accurate prediction of the number of genes sequenced. Non-coding DNA (e.g. UTR - UnTranslated Region or intron readthroughs) sequenced from multiple haplotypes is not easy to assemble due to high levels of heterozygosity caused by



Species	*D. melan.*	*B. anynana*	*M. cinxia*	*M. sexta*	*M. sexta*	*C. tremulae*	*E. aurinia*	*H. erato*	*H. melpomene*	*P. dardanus*
Tissue	various	various	whole	hemocyte	midgut	midgut	whole	wing disc	wing disc	wing disc
Technology	Sanger	Sanger	2x GS20	GS20	GSFLX	GSFLX	GSFLX	GSFLX	2.5 x GSFLX	GSFLX Titan.
Before Preprocess	820,813	101,606	595,941	95,458	431,753	277,780	195,401	313,651	879,750	940,005
After Preprocess	745,124	96,977	536,696	82,267	379,189	252,828	173,351	246,191	735,162	531,702

Figure 4
Boxplot of read length before and after pre-processing for each dataset, showing 25% and 75% intervals, the horizontal bar shows the median, the diamond shows the mean, whiskers encompass entire data range. Such information offers an overall picture of a sequencing run's quality.

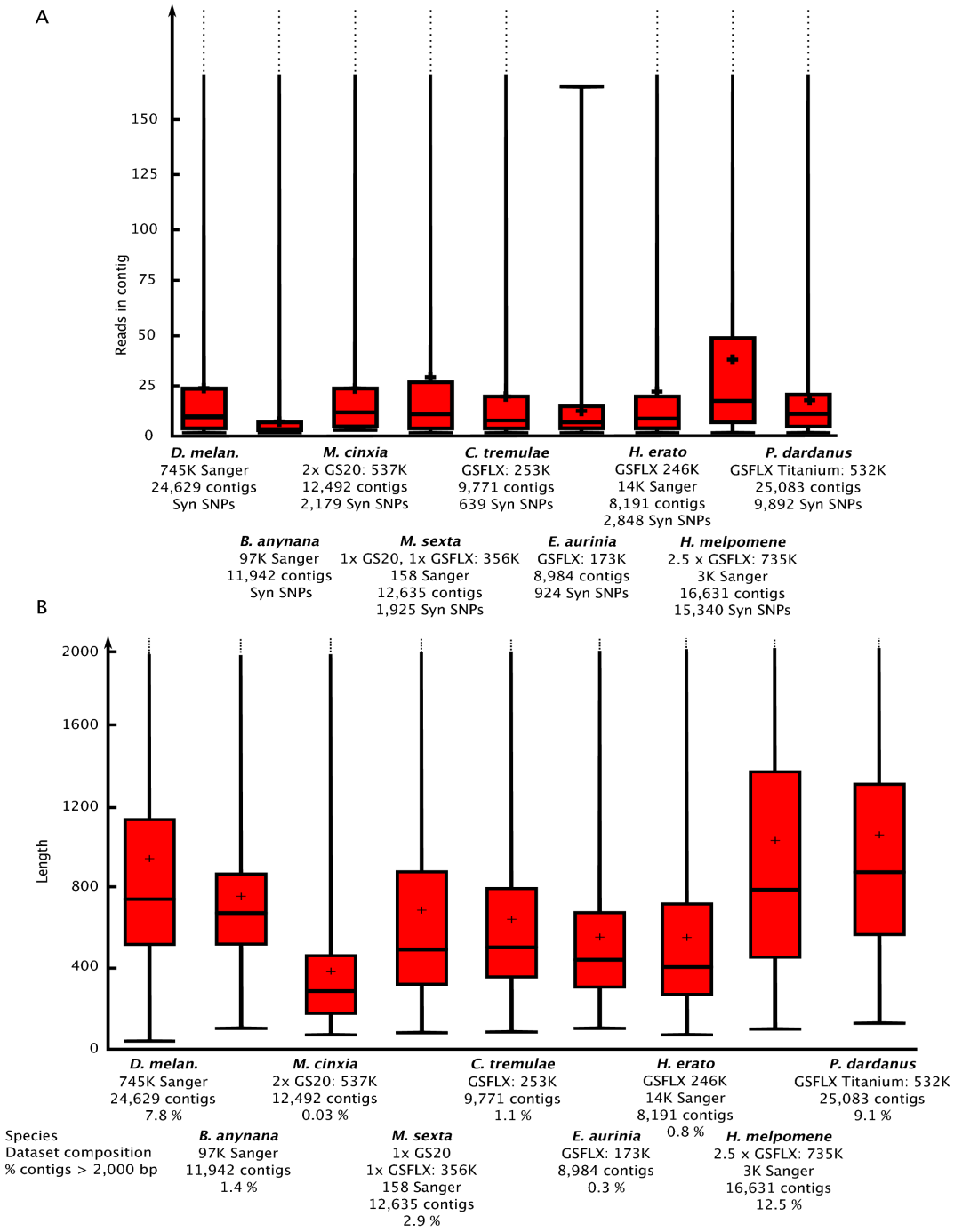


Figure 5

Boxplot of (A) number of reads assembled in each contig and (B) contig length for each dataset, showing 25% and 75% intervals, the horizontal bar shows the median, the plus sign shows the mean, whiskers encompass entire data range. Sanger technology has been considered a cleaner technique despite a higher cost but the B. anynana dataset (ca 97K sequences) performs poorly when compared to GSFLX. The earlier GS20 technology is significantly inferior and of limited use in transcriptome sequencing.

BMC Bioinformatics 2009, **10**:447 http://www.biomedcentral.com/1471-2105/10/447

the fact that the majority of the UTR is evolving without constraints. Upon investigation, this seems to be the major cause of contig inflation and the platform can use two methods to alleviate the issue by filtering in favor of regions likely to be coding. One method is used in Figure 5 but we may have the unfortunate effect of removing small contigs containing novel proteins which are small in size and low in expression. The other method - available only to users with a dataset from related species - is to use a tBLASTx approach (via analyse_assembly.pl) to complement the BLASTx approach of the reference proteome. Novel proteins, even if evolving rapidly, are expected to show significant similarity on the amino acid level between the two species. The platform allows one to extract the contig regions matching this coding fraction and thus have a dataset known to be coding. The two approaches are not mutually exclusive but can be complementary: the trim_assembly.pl is highly customizable regarding similarity and abundance levels whereas the tBLASTx approach will not tackle the issue of redundancy (i.e. two contigs originating from one locus).

We can, therefore, conclude that a better assembly benchmark is the identification of proteins from a reference proteome and the portion of the assembly identified as coding (CDS; CoDing Sequence). The quality of a sequencing experiment can thus be evaluated by extracting the CDS fraction and then calculating the proportion of reads contributing to this portion of the assembly (e.g. by doing a BLAST similarity search). In our dataset, we find that the *Drosophila* dataset covers only 70% of the *Drosophila* proteome and only 56% of the assembly is defined as coding (Figure 6). From the BioMart.org website we calculated that the proportion of non-UTR in annotated *D. melanogaster* genes is only 80%. The trend of having a significantly lower CDS proportion than expected is maintained across the data when a BLASTx versus a reference proteome is used. One reason can be purely technical: in a dataset originating from a library with large number of haplotypes, there is a contig inflation when the UTR is included for assembly. This effect is more likely to be present in non-model species where isogenic lines or sufficient levels of inbreeding are unattainable. The second technical issue, especially in NGS datasets, is that due to the fragmentation step involved, there can be a preference for sequencing of the transcript ends and therefore UTR [52].

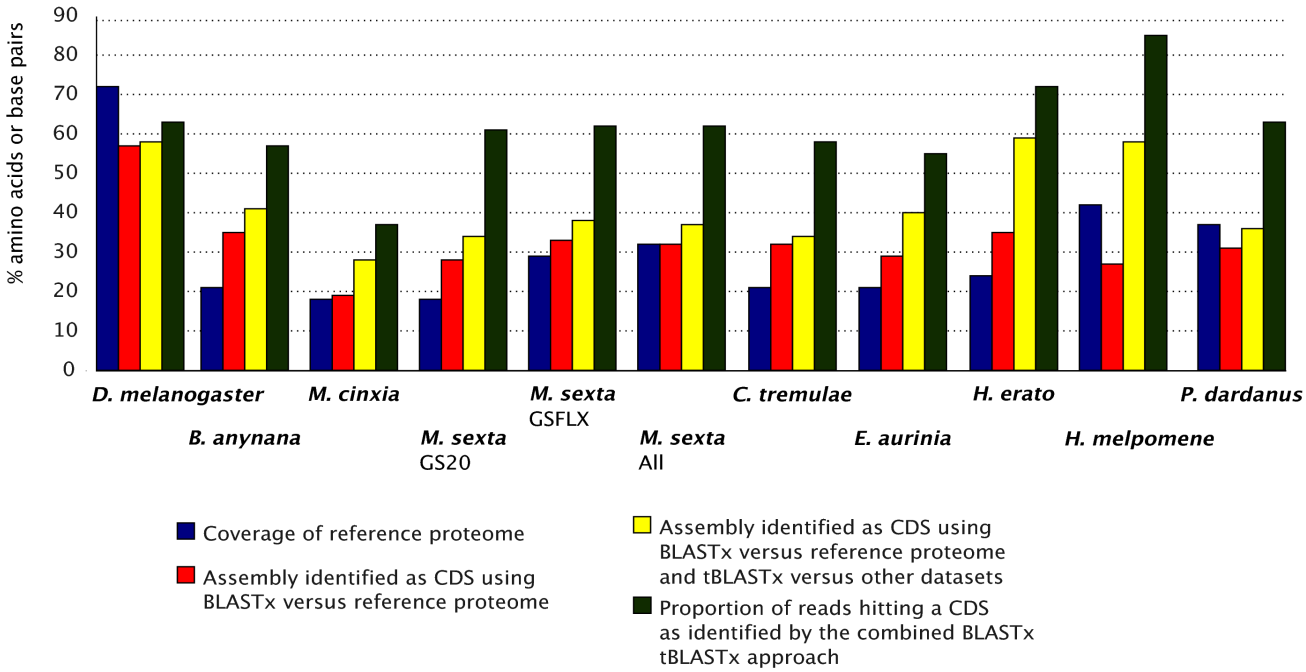


Figure 6
Comparison of the number of genes and proteins identified using different 454 based sequencing technologies (GS20, GSFLX and GSFLX-Titanium). For each dataset, the accuracy of the results depends on how similar the target and reference transcriptomes are and the improvement with tBLASTx is an indication of novel protein data supported by at least two species. Such cases warrant a more thorough investigation and can result in the determination of taxon specific- or rapidly evolving genes. The proportion of reads from the sequencer (after pre-processing) which are part of these coding regions is also shown. This can guide future project designs which wish to aim to alter the representation of non-coding in the sequenced sample.

BMC Bioinformatics 2009, **10**:447 http://www.biomedcentral.com/1471-2105/10/447

Regardless whether a measure against redundancy is used, if multiple datasets are available, one can explore whether a reference proteome is useful and whether using a phylogenetic framework assists in gene-finding. Due to the taxon focus and species richness in our data, we expect that mutual tBLASTx searches among the assembled NGS datasets (excluding the species used a query) as reference databases, will identify additional proteins which may be absent from the reference proteome, or sufficiently diverged to be missed by comparison to it. Because the *B. anynana*, *M. cinxia*, *E. aurinia*, *P. dardanus*, *H. erato* and *H. melpomene* datasets originate from butterflies with the latter two being from the same genus (*Heliconius*) and the *M. sexta* datasets originate from a moth in the same superfamily as the reference proteome *B. mori* [53] we expect and find that the butterflies will show a significantly higher improvement with tBLASTx than *M. sexta*. Oddly, we also find a large improvement in *C. tremulae*, a beetle which uses the *Tribolium castaneum* as a reference but are in different superfamilies. This improvement is unlikely to be due to a poorer annotation in *Tribolium* versus *Bombyx* as the reference annotation proportions are relatively similar in all non-model species datasets. It is not unlikely that *Tribolium* is not a good model for *C. tremulae* especially if one begins to consider the difference in their ecology. The other, not mutually exclusive, possibility is that there is a significant degree of rapidly evolving proteins in these non-model species. Nevertheless, via the tBLASTx approach, each of the *Heliconius* data has now a CDS proportion comparable to the *Drosophila* gold standard. Indeed, by also counting the number of sequence reads belonging to the estimated CDS (Figure 7) we are able to see that the *Heliconius* datasets have also a higher proportion of reads belonging to the CDS. This observation provides some evidence that contig inflation in this case is more likely to be driven by unassembled UTR rather than due to fragmentation of the transcript ends.

One other point of note is relating to the procedure of choosing a cDNA generation protocol. All NGS datasets compared here use the SMART technology to produce cDNA apart from the GS20 dataset of *M. sexta* which was produced using GC-rich random primers. We do not know why the number of reads was much lower than the GS20 dataset from *M. cinxia* but a significantly higher proportion of the reads matches a predicted CDS which translates to a better return to projects aimed at gene-hunting. We cannot be sure why this occurs: it could be due to a high number of low quality reads in *M. cinxia*, it could be due to a slightly better protein identification based on the closely related *B. mori* but it is logical to expect that, in species known to have a GC enriched coding sequences, the primer protocol would enrich for regions with high GC content and therefore more likely target coding regions.

Finally, researchers often wish to know how deep one should sequence in order to sequence the complete transcriptome in their sample. This is important in planning

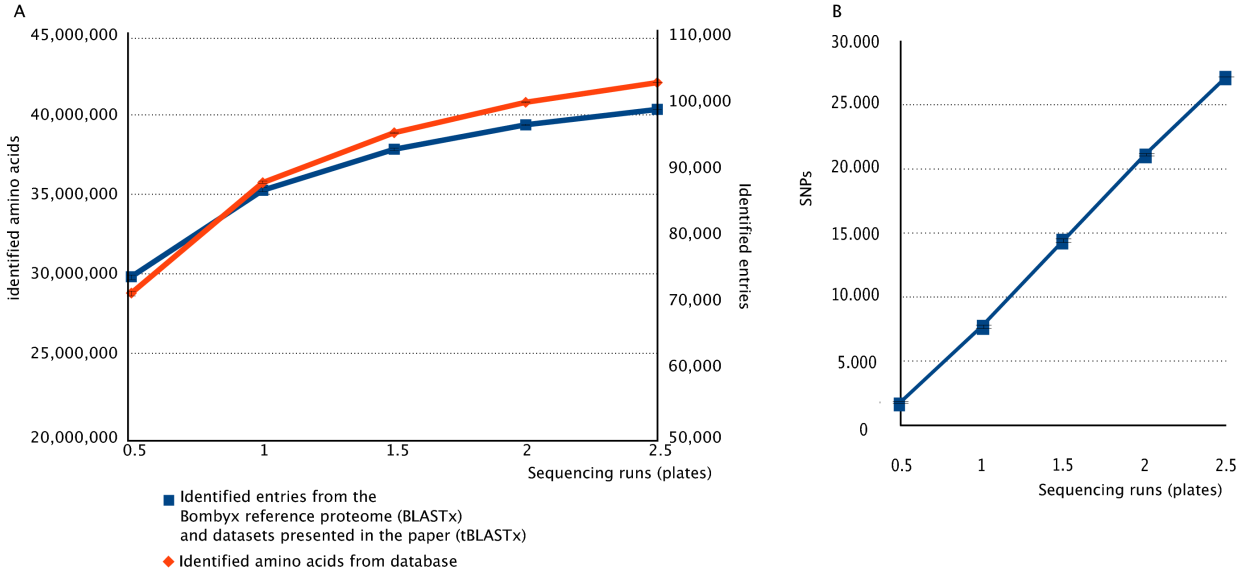


Figure 7
Saturation curve of 454 GSFLX sequencing using the *H. melpomene* dataset. The error bars show the min/max of each data point as verified with 5 independent pseudo-samples. (A) Researchers can obtain a substantial number of genes with data from one half-plate with saturation for the transcriptome of this sample near the 2.5 plates. (B) SNP marker identification is linear in this dataset with an average of 1,757 high quality SNPs identified in one half-plate.

BMC Bioinformatics 2009, **10**:447 http://www.biomedcentral.com/1471-2105/10/447

non-model species transcriptomes projects. As the *H. melpomene* 454 data originate from 2.5 full-plates (i.e. 5 half-plates) and the cDNA is harvested from a single tissue (wing discs) we used 20 pseudo-assemblies to investigate the effect of deeper sequencing. As Figure 7 shows, most transcripts for that particular tissue were identified after 1.5 half plates were sequenced as the exponential curve is approaching a plateau (Figure 7A). With 1 half-plate, however, 74.5% of the plateau value for proteins identified was attained, showing that even shallow sequencing of a non-model species is highly worthwhile. For SNP detection, the function is not exponential (Figure 7B): with each subsequent run the number of high quality SNP markers increased linearly.

Discussion

There are several advantages of *est2assembly* over other platforms for processing EST raw data (e.g. [54,55]). First, preprocessing of raw sequence is essential and our platform offers a standard method for consistently accomplishing this for hundreds of thousands of sequences with straightforward user customization. Second, parameterizing an assembler is a tedious process and our platform is the only one which automates many of the routines. Third, annotation of an assembly with *est2assembly* can be readily standardized and automated for processing large numbers of datasets with minimum investment in time. Deciding on the optimal assembly is a subjective process and depends on the project but by providing the means to explore the parameter space allows for a standardization of an approach which is often ad-hoc. We calculate the BLAST-based index using two approaches and can determine the number of unique proteins found, actual proportion of amino acids found and obtain an estimate of the assembly proportion that is actually coding. In this case, as more datasets are published, we can benchmark laboratory protocols and sequencing technologies involved in acquiring full length genes. Importantly, the rich log output guides the wet-lab biologist who generated the data to perform in-depth investigations and hold a better understanding of their project. With *est2assembly*, we have not aimed to produce a 'black box' but a program which gives feedback to the user as to the quality and characteristics of the different assemblies achieved with their data. We showed that analysis of an assembly can give important insights to the technologies and protocols employed to acquire a transcriptome. Future work can focus on including more annotation modules and developing a Java/JDBC-driven Graphical User Interface (GUI) and relational database to allow molecular biologists with no computing knowledge to supervise the data analysis.

Shallow sequencing EST projects are becoming a goldmine for biologists working on non-model species and are often used for both gene or SNP discovery but until now no software exists to link the SNP to both an assembly and the codon it may be part of. The *est2assembly* platform allows for the classification and identification of SNPs which may be under selection or point to a misalignment. Such data are important in manual curation of an assembly and lacking from any other software. We can also obtain coding synonymous SNPs for which a PCR primer is straightforward to design but are under low levels of selection. Non-coding markers are also useful to researchers who wish to investigate selection in non-coding DNA.

Special considerations, however, have to apply to projects working on non-model species, especially when datasets are restricted for financial or biological reasons. Often the design of the experiment is not conceived with full knowledge of a technology's capabilities and limitations. Here we show that different technologies and lab protocols differ in their ability to produce an assembly and project design plays an important role. At times, but not always, such project design bottlenecks can be overcome. For example, assemblers treat the common issue problem of inflated contig number caused by non-optimal alignments by assuming that they are based solely on sequencing errors or repeats. The result is that fewer genes are discovered in non-model species. The issue is confounded because researchers undertake a transcriptome project for different reasons. Even though in gene discovery project the norm is to sequence a cDNA library from specific tissues and with a limited number of haplotypes, this is rarely the case in most EST projects of non-model species. Researchers often utilize EST projects as both a gene-finding project and a SNP discovery protocol and therefore are tempted to include a high number of out-bred individuals. It should be noted that both Newbler and MIRA are not clustering algorithms but assemblers and therefore their main aim is not to identify alternative splicing events or cope with a high degree of heterozygosity in the sequenced sample. There are methods to alleviate the problem such as including a final clustering step (e.g. miraEST [56]; CLOBB [57]; or CAP3 [58]. Our platform does not yet contain such a clustering step as the levels of heterozygosity can vary and a supervised algorithm we are developing as a future module may provide a more optimal solution. Such an algorithm would be tailored (i.e. trained) for each transcriptome project and make assignment of supercontigs (e.g. the merging of alternative splicing events and non-coding regions belonging to the same locus) more robust. Another issue which cannot be resolved using bioinformatics is the quality of material used for cDNA preparation. Even though we cannot be certain regarding the cause, the *M. cinxtia* and *E. aurinia* datasets argue against a whole animal approach in constructing the cDNA library. Such cases have been shown to be problematic in enzymatic reactions due to PCR-inhib-

iting pigments [59]. The inclusion of micro organisms in the digestive tract or the outer body can also result in acquiring contaminating sequence from another species. The later cases can be investigated bioinformatically [60,61] so as to prevent generating an erroneous transcriptome survey.

Conclusion

In conclusion, the modern transcriptome sequencing approaches are very powerful and cost effective but they still yield partial transcriptomes. In the future, however, our single most important limitation will not be raw transcriptomic or genomic data. We have shown that the ability to accurately annotate an assembly depends on using a correct reference proteome or utilize phylogenetic framework. Further, comparison to an appropriate reference proteome is invaluable in choosing among different assemblies, yet such proteomes are themselves incomplete. A concerted annotation effort based on transcriptome sequencing from a diverse phylogenetic collection is required, which will accelerate the filling of proteome space beyond the limited set of model organisms that currently occupy it. With the NGS capabilities, it is obvious that such an effort should make full use of transcriptomic data in order but we still lack the necessary infrastructure. The *est2assembly* software, however, has enabled the development of such infrastructure, an alpha stage preview of which is available at http://www.insectacentral.org.

Availability and requirements

Project name: *est2assembly*

Project home page: http://code.google.com/p/est2assembly/

Operating system: Linux

Programming language: Perl

Dependencies on proprietary software: None

Other requirements: NCBI-BLAST, EMBOSS toolkit, BioPerl dev-branch, PostgreSQL, MIRA (included), sff_extract (included)

Optional programs (included): annot8r, prot4EST

License: General Public License version 3

Abbreviations

BAC: Bacterial Artificial Chromosome; CDS: CoDing Sequence; EBI: European Bioinformatics Institute; EST: Expressed Sequence Tag; GFF: General Feature Format; GMOD: Generic Model Organism Database; GPL: General Public License; GUI: Graphical User Interface; NGS: Next Generation Sequencing (technology); ORF: Open Reading Frame; PAS: PolyAdenylation Signal; SNP: Single Nucleotide Polymorphism; UTR: UnTranslated Region.

Authors' contributions

AP conceived, designed and performed the study; analyzed and interpreted data; coded the software and drafted the manuscript. RS co-authored the GFF writing software and the GBrowse schema. RHfC and DGH drafted the manuscript, financed and provided infrastructure for the study. All authors approved the final version of the manuscript.

Acknowledgements

We would like to thank Karl Gordon (CSIRO) for helping with end-user testing, two anonymous referees for improving the manuscript and the following for making pre-publication data available: Chris Jiggins and his laboratory (Univ. of Cambridge), Owen McMillan and his laboratory (State Univ. of N. Carolina), Yannick Pauchet and Iva Fuková (Univ. of Exeter). Further, Bastien Chevreux provided development versions of MIRA and excellent support, Jose Blanca provided sff_extract, James Wasmuth provided support for prot4EST, Ralf Schmid for annot8r, Derek Huntley for SEAN and Steffi Gebauer-Jung for TrimbyWindow. David Clements and Scott Cain helped with Chado and GBrowse. We also thank the TU-Dresden Deimos PC-Farm for computational support. The authors report no conflicting interests. AP was supported by the Max Planck Gesellschaft and the European Union Research Network GAMEXP; DGH was supported by the Max Planck Gesellschaft; RHfC was supported by the European Union Research Network EMBEK1.

References

1. Van Straalen NM, Roelofs D: **An introduction to ecological genomics.** Oxford: Oxford University Press; 2006.
2. Heckel DG, Gahan LJ, Daly JC, Trowell S: **A genomic approach to understanding *Heliothis* and *Helicoverpa* resistance to chemical and biological insecticides.** *Philos Trans R Soc Lond B Biol Sci* 1998, **353:**1713-1722.
3. Brakefield PM, Gates J, Keys D, Kesbeke F, Wijngaarden PJ, Monteiro A, French V, Carroll SB: **Development, plasticity and evolution of butterfly eyespot patterns.** *Nature* 1996, **384:**236-242.
4. Rausher MD: **Natural selection and the evolution of plant insect interactions.** In *Insect chemical ecology: an evolutionary approach* Edited by: Rausher MD, Isman MB. New York: Chapman & Hall; 1992:20-88.
5. Ewing B, Green P: **Analysis of expressed sequence tags indicates 35,000 human genes.** *Nat Genet* 2000, **25:**232-234.
6. Rudd S: **Expressed sequence tags: alternative or complement to whole genome sequences?** *Trends Plant Sci* 2003, **8:**321-329.
7. Boguski MS, Lowe TMJ, Tolstoshev CM: **dbEST-- database for "expressed sequence tags".** *Nat Genet* 1993, **4:**332-333.
8. Beldade P, McMillan WO, Papanicolaou A: **Butterfly genomics eclosing.** *Heredity* 2008, **100:**150-157.
9. Mita K, Morimyo M, Okano K, Koike Y, Nohata J, Kawasaki H, Kadono-Okuda K, Yamamoto K, Suzuki MG, Shimada T: **The construction of an EST database for Bombyx mori and its application.** *Proc Natl Acad Sci* 2003, **100:**14121-14126.
10. Mita K, Kasahara M, Sasaki S, Nagayasu Y, Yamada T, Kanamori H, Namiki N, Kitagawa M, Yamashita H, Yasukochi Y: **The genome sequence of silkworm, Bombyx mori.** *DNA Res* 2004, **11:**27-35.
11. Papanicolaou A, Gebauer-Jung S, Blaxter ML, McMillan WO, Jiggins CD: **ButterflyBase: a platform for lepidopteran genomics.** *Nucleic Acids Res* 2008, **36:**D582-587.
12. Bouck A, Vision T: **The molecular ecologist's guide to expressed sequence tags.** *Mol Ecol* 2007, **16:**907-924.
13. Thomson RC, Shedlock AM, Edwards SV, Shaffer HB: **Developing markers for multilocus phylogenetics in non-model organ-**

isms: A test case with turtles. *Mol Phylogenet Evol* 2008, **49:**514-525.
14. Papanicolaou A, Joron M, McMillan WO, Blaxter ML, Jiggins CD: **Genomic tools and cDNA derived markers for butterflies.** *Mol Ecol* 2005, **14:**2883-2897.
15. Schuster SC: **Next-generation sequencing transforms today's biology.** *Nat Methods* 2008, **5:**16-18.
16. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S: **The Generic Genome Browser: A Building Block for a Model Organism System Database.** *Genome Res* 2002, **12:**1599-1610.
17. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H: **The Bioperl toolkit: Perl modules for the life sciences.** *Genome Res* 2002, **12:**1611-1618.
18. Mungall CJ, Emmert DB: **A Chado case study: an ontology-based modular schema for representing genome-associated biological information.** *Bioinformatics* 2007, **23:**i337-346.
19. Chevreux B, Wetter T, Suhai S: **Genome sequence assembly using trace signals and additional sequence information.** *Proc German Conf Bioinformatics* 1999, **99:**45-56.
20. **SFF extract** [http://bioinf.comav.upv.es/sff_extract/]
21. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite.** *Trends Genet* 2000, **16:**276-277.
22. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
23. Ning Z, Cox AJ, Mullikin JC: **SSAHA: A Fast Search Method for Large DNA Databases.** *Genome Res* 2001, **11:**1725-1729.
24. **RepeatMasker** [http://www.repeatmasker.org]
25. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Repbase Update, a database of eukaryotic repetitive elements.** *Cytogenet Genome Res* 2005, **110:**462-467.
26. Wasmuth JD, Blaxter ML: **Prot4EST: Translating Expressed Sequence Tags from neglected genomes.** *BMC Bioinformatics* 2004, **5:**187.
27. Pearson WR, Wood T, Zhang Z, Miller W: **Comparison of DNA sequences with protein sequences.** *Genomics* 1997, **46:**24-36.
28. Schmid R, Blaxter ML: **annot8r: GO, EC and KEGG annotation of EST datasets.** *BMC Bioinformatics* 2008, **9:**180.
29. **Phred, Phrap, and Consed** [http://www.phrap.com]
30. Ji G, Zheng J, Shen Y, Wu X, Jiang R, Lin Y, Loke J, Davis K, Reese G, Li Q: **Predictive modeling of plant messenger RNA polyadenylation sites.** *BMC Bioinformatics* 2007, **8:**43.
31. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT: **Gene Ontology: tool for the unification of biology.** *Nat Genet* 2000, **25:**25-29.
32. Bairoch A: **The ENZYME database in 2000.** *Nucleic Acids Res* 2000, **28:**304-305.
33. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Res* 2000, **28:**27-30.
34. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Res* 2001, **29:**308-311.
35. Huntley D, Baldo A, Johr S, Sergot M: **SEAN: SNP prediction and display program utilizing EST sequence clusters.** *Bioinformatics* 2006, **22:**495.
36. Zou Z, Najar F, Wang Y, Roe B, Jiang H: **Pyrosequence analysis of expressed sequence tags for *Manduca sexta* hemolymph proteins involved in immune responses.** *Insect Biochem Mol Biol* 2008, **38:**677-682.
37. Pauchet Y, Wilkinson P, van Munster M, Augustin S, Pauron D, Ffrench-Constant RH: **Pyrosequencing of the midgut transcriptome of the poplar leaf beetle Chrysomela tremulae reveals new gene families in Coleoptera.** *Insect Biochem Mol Biol* 2009, **39:**403-13.
38. Pauchet Y, Wilkinson P, Vogel H, Nelson DR, Reynolds SE, Heckel DG, ffrench-Constant RH: **Pyrosequencing *Manduca sexta* larval midgut transcriptome: messages for digestion, detoxification and defence.** *Insect Mol Biol* in press.
39. Ferguson L, Lee SF, Chamberlain N, Nadea N, Joron M, Baxter S, Wilkinson P, Papanicolaou A, Kumar S, Thuan-Jin Clark R, Davidson C, Glithero R, Beasle H, Vogel H, Ffrench-Constant R H, Jiggins CD: **Characterization of a hotspot for mimicry: assembly of a butterfly wing transcriptome to genomic sequence at the HmYb/Sb locus.** *Mol Ecol* in press.
40. Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, Marden JH: **Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing.** *Mol Ecol* 2008, **17:**1636-47.
41. Beldade P, Saenko SV, Pul N, Long AD: **A Gene-Based Linkage Map for *Bicyclus anynana* Butterflies Allows for a Comprehensive Analysis of Synteny with the Lepidopteran Reference Genome.** *PLoS Genet* 2009, **5:**e1000366.
42. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2007, **35:**D61-65.
43. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M: **The universal protein resource (UniProt).** *Nucleic Acids Res* 2005, **33:**D154-159.
44. Drysdale RA, Crosby MA: **FlyBase: genes and gene models.** *Nucleic Acids Res* 2005, **33:**D390-395.
45. Lawson D, Arensburger P, Atkinson P, Besansky NJ, Bruggner RV, Butler R, Campbell KS, Christophides GK, Christley S, Dialynas E: **VectorBase: a home for invertebrate vectors of human pathogens.** *Nucleic Acids Res* 2007, **35:**D503-505.
46. Lawson D, Arensburger P, Atkinson P, Besansky NJ, Bruggner RV, Butler R, Campbell KS, Christophides GK, Christley S, Dialynas E: **VectorBase: a data resource for invertebrate vector genomics.** *Nucleic Acids Res* 2009, **37:**D583-587.
47. Solignac M, Zhang L, Mougel F, Li B, Vautrin D, Monnerot M, Cornuet JM, Worley KC, Weinstock GM, Gibbs RA: **The genome of Apis mellifera: dialog between linkage mapping and sequence assembly.** *Genome Biol* 2007, **8:**403.
48. Wang J, Xia Q, He X, Dai M, Ruan J, Chen J, Yu G, Yuan H, Hu Y, Li R: **SilkDB: a knowledgebase for silkworm biology and genomics.** *Nucleic Acids Res* 2005, **33:**D399.
49. Wang L, Wang S, Li Y, Paradesi MSR, Brown SJ: **BeetleBase: the model organism database for *Tribolium castaneum*.** *Nucleic Acids Res* 2007, **35:**D476-479.
50. Yamamoto K, Narukawa J, Kadono-Okuda K, Nohata J, Suetsugu Y, Sasanuma M, Sasanuma S, Mita K, Minami H, Shimomura M: **Silkworm genome analysis: Construction of an integrated genome database, KAIKObase.** *Seikagaku* 2006, **A12627:**78.
51. Li W, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.** *Bioinformatics* 2006, **22:**1658-9.
52. Harismendy O, Frazer K: **Method for improving sequence coverage uniformity of targeted genomic intervals amplified by LR-PCR using Illumina GA sequencing-by-synthesis technology.** *BioTechniques* 2009, **46:**229.
53. Goldsmith MR, Shimada T, Abe H: **The genetics and genomics of the silkworm, *Bombyx mori*.** *Annu Rev Entomol* 2005, **50:**71-100.
54. Parkinson J, Anthony A, Wasmuth J, Schmid R, Hedley A, Blaxter M: **PartiGene - constructing partial genomes.** *Bioinformatics* 2004, **20:**1398-1404.
55. Paquola AC, Nishyiama Jr MY, Reis EM, da Silva AM, Verjovski-Almeida S: **ESTWeb: bioinformatics services for EST sequencing projects.** *Bioinformatics* 2003, **19:**1587-1587.
56. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WEG, Wetter T, Suhai S: **Using the miraEST Assembler for Reliable and Automated mRNA Transcript Assembly and SNP Detection in Sequenced ESTs.** *Genome Res* 2004, **14:**1147-1159.
57. Parkinson J, Guiliano DB, Blaxter M: **Making sense of EST sequences by CLOBBing them.** *BMC Bioinformatics* 2002, **3:**31.
58. Huang X, Madan A: **CAP3: A DNA sequence assembly program.** *Genome Res* 1999, **9:**868-877.
59. Bextine B, Tuan S, Shaikh H, Blua M, Miller TA: **Evaluation of Methods for Extracting *Xylella fastidiosa* DNA from the Glassy-Winged Sharpshooter.** *J Econ Entomol* 2004, **97:**757-763.
60. Friedel CC, Jahn KHV, Sommer S, Rudd S, Mewes HW, Tetko IV: **Support vector machines for separation of mixed plant-pathogen EST collections based on codon usage.** *Bioinformatics* 2005, **21:**1383-1388.
61. Emmersen J, Rudd S, Mewes HW, Tetko IV: **Separation of sequences from host-pathogen interface using triplet nucleotide frequencies.** *Fungal Genet Biol* 2007, **44:**231-241.

Chapter 3 - ButterflyBase: a platform for lepidopteran genomics

This Chapter was published in 2008 and provided the first dedicated transcriptome database which is widely used by the lepidopteran community (it was published in January 2008 and has been cited at least 23 times since then; source: ISI Web of Knowledge accessed 03 October 2010). The deployment made use of existing software (PartiGene) but the provision of reference sequence generated from transcriptome data for an entire taxon was innovative. This proof of concept paper showed how reference transcriptome research can benefit the EEFG field even when reference genomic sequence is lacking.

Citation

Papanicolaou, A. et al., 2008. ButterflyBase: a platform for lepidopteran genomics. Nucleic acids research, 36, D582-7.

Nucleic Acids Research Advance Access published October 12, 2007

Nucleic Acids Research, 2007, ***1–6***
doi:10.1093/nar/gkm853

ButterflyBase: a platform for lepidopteran genomics

Alexie Papanicolaou[1,2,*], Steffi Gebauer-Jung[2], Mark L. Blaxter[1], W. Owen McMillan[3] and Chris D. Jiggins[1,4]

[1]Institute for Evolutionary Biology, University of Edinburgh, King's Buildings, EH9 3JT, UK, [2]Max Planck Institute for Chemical Ecology, Jena, 07745, Germany, [3]Department of Genetics, North Caroline State University, Raleigh, NC 27695-7614, USA and [4]Department of Zoology, University of Cambridge, Downing Street, CB2 3EJ, UK

Received August 15, 2007; Revised September 25, 2007; Accepted September 26, 2007

ABSTRACT

With over 100 000 species and a large community of evolutionary biologists, population ecologists, pest biologists and genome researchers, the Lepidoptera are an important insect group. Genomic resources [expressed sequence tags (ESTs), genome sequence, genetic and physical maps, proteomic and microarray datasets] are growing, but there has up to now been no single access and analysis portal for this group. Here we present ButterflyBase (http://www.butterflybase.org), a unified resource for lepidopteran genomics. A total of 273 077 ESTs from more than 30 different species have been clustered to generate stable unigene sets, and robust protein translations derived from each unigene cluster. Clusters and their protein translations are annotated with BLAST-based similarity, gene ontology (GO), enzyme classification (EC) and Kyoto encyclopaedia of genes and genomes (KEGG) terms, and are also searchable using similarity tools such as BLAST and MS-BLAST. The database supports many needs of the lepidopteran research community, including molecular marker development, orthologue prediction for deep phylogenetics, and detection of rapidly evolving proteins likely involved in host–pathogen or other evolutionary processes. ButterflyBase is expanding to include additional genomic sequence, ecological and mapping data for key species.

INTRODUCTION AND MOTIVATION

The Lepidoptera (butterflies and moths) are remarkably diverse containing more than 100 000 described species. There is a long tradition of research and a number of disciplines use lepidopteran models to investigate fundamental biological phenomena including development and gene regulation, population genetic processes (gene flow, colonization and extinction), adaptation and morphological innovation, speciation and co-evolutionary processes such as host–plant and insect–parasite interactions. As a result, there is a wealth of ecological and genetic knowledge for Lepidoptera.

The silkworm *Bombyx mori* is a model for insect physiology and molecular biology, as well as being an important crop animal. Currently, two whole genome shotgun sequence assemblies are publicly available (1,2) and a joint genome assembly by the Chinese and Japanese teams is expected within 2007. The genomic sequence data are anchored by a number of bacterial artificial chromosome (BAC) libraries, high-density linkage maps of sequence tag sites (STS), cDNA and microsatellite (simple sequence repeats, SSR) markers (3–6) as well as cytogenetic studies (7) which provide a chromosomal framework for genome assembly. Thus the chromosomal framework for genome assembly is in place and as the annotation of the *B. mori* genome progresses, it will facilitate comparative analysis of other species with less complete genomic information (8).

In addition to genomic resources in *Bombyx*, there is increasing amount of EST data for a growing number of Lepidoptera species. Large to moderate-sized EST datasets are becoming easier and less expensive to produce and can be powerful source of markers for comparative mapping, population genetic analysis and studies of adaptive evolution (9). For example, there are large public genomic datasets for the moth pest *Spodoptera frugiperda*, and the butterflies *Bicyclus anynana*, *Heliconius melpomene* and *Heliconius erato*. The generation of sequences for these and other species has led to the discovery that around half of the sequenced genes in Lepidoptera have little or no sequence similarity to proteins from other taxa (8). Species-specific public databases are available for these taxa, but vary widely in accessibility and format (10–13). What is lacking is a central platform for accessing lepidopteran data and more importantly for conducting comparative between species analyses.

To allow the community to benefit from the comparative genomic data available in Lepidoptera, we developed

*To whom correspondence should be addressed. Tel: +493641571561; Fax: +493641571502; Email: alexie@butterflybase.org

an online database and annotation platform, called ButterflyBase. It is available at http://www.butterflybase.org. ButterflyBase is a comparative gene-focused database for all Lepidoptera. ButterflyBase brings together, in a single site, sequence information for all lepidopterans including *B. mori*. ButterflyBase was designed to extend the utility of the publicly available expressed sequence tag (EST) datasets using clustering and protein prediction software, and to provide high-quality annotation for data mining and exploitation, all through a simple and intuitive user interface. With this short article, we hope to introduce users to the utility of the database. Further information regarding technical details can be obtained on request or by browsing the dataset download page.

METHODOLOGY

Datasets

ESTs and full-length cDNA sequences were obtained from public depositions in the EMBL/GenBank/DDBJ database, and clustered using a modified version of the PartiGene suite (14). When the original sequencer chromatograms were available (*H. erato* and *H. melpomene*) we processed them with trace2dbest (14). All other data were pre-processed to remove vector contamination, poly(A) tails and sequences smaller than 150 bp. For some cDNA libraries (where sequence quality was poor), further trimming was performed using a customized version of est_trimmer.pl [provided by Thomas Thiel through the MISA program (15)]. SSR prediction was performed using MISA (15), single nucleotide polymorphisms (SNPs) were predicted using SEAN (16) and databased using custom Perl scripts. A SEAN Java viewer is available as a modified applet, provided by the SEAN author. The methodology of SEAN does not rely on quality information and therefore can be used with our datasets. Instead, it only marks putative SNPs if a single nucleotide change is present in at least two members of the EST cluster and there are no other nucleotide inconsistencies 15 bp up- and downstream of the putative SNP.

PartiGene (14) uses megablast and the CLOBB approach to cluster EST sequences into groups putatively derived from the same mRNA molecule (17). These clusters are subsequently aligned using Phrap (with the *forcelevel* option set to maximum) (18, Green,P., unpublished software). Sequenced organisms in Lepidoptera are often outbred and may, therefore, exhibit substantial allelic variation. Essentially, the presence of low quality, multiple SNPs, sequencing errors, alternative splicing or short indels may allow megablast to generate a cluster of highly similar sequences which is not subsequently aligned by Phrap, thus leading to some clusters containing more than one contig.

ButterflyBase uses a two-letter code to signify the species ID and a third letter to signify molecule type (P for protein, C for nucleotide cluster (or unigene) and in the future B for BAC clone). Each cluster of ESTs and cDNAs has a unique numerical ID, which is stable when additional sequences are added to the dataset. When there is more than one contig per cluster these are indicated by a trailing number. Thus HEC00123_1 is the first contig of a nucleotide cluster from *H. erato* and its protein translation is HEP00123_1. Cluster identifiers are conserved as more sequences are added.

Protein prediction

The protein predictions are ButterflyBase's strongest asset. We use, prot4EST, a protein prediction tool developed specifically for EST data (19). Briefly, this program utilizes a four-tier methodology: first, similarity to known proteins is used in order to detect the open reading frame (ORF) and correct for any potential sequencing errors [using the high-scoring segment pair (HSP) tiling approach], if that fails (e.g. for novel or Lepidoptera-specific genes) ESTSCAN is utilized (20) and if that fails too then DECODER (21) and finally the longest ORF from the six-frame translation. As prior training data (codon usage tables and base composition estimates) for probabilistic prediction of ORFs were not available for many lepidopteran species, we utilized data derived from high-scoring BLAST matches to populate species-specific parameter sets.

Database schema and dataset annotation

The database is driven by PostgreSQL with a customized version of the PartiGene schema. The central entity is a mRNA sequence cluster. Each cluster is annotated with a number of facilities. The most frequently accessed are pre-computed BLAST similarity searches versus a variety of databases: Uniref100; a collection of possible contaminants (e.g. fungi, viruses, bacteria, molecular biology vectors) and phylogenetically selected, nested databases. We chose a number of such databases including *B. mori* nucleotides and proteins; Lepidoptera nucleotides without *B. mori*; proteins from released Arthropoda genomes; Arthropoda sequences without those genomes or Lepidoptera. All BLAST searches have an *E*-value cutoff of $1E{-}4$. Furthermore, predictions enhance the utility of the consensus: a robust protein translation as well as SSR and SNP predictions are currently offered. The protein predictions in turn are annotated with enzyme classification (EC), gene ontology (GO) and Kyoto encyclopaedia of genes and genomes (KEGG) terms. These latter annotations are derived from BLAST searches of annotated protein databases using the annot8r tool (Schmid,R. and Blaxter,M., unpublished software), and a cut-off *E*-value of $1E{-}8$. Furthermore, ButterflyBase provides domain annotations from InterProScan (22) and basic protein statistics to facilitate downstream proteomic and biochemical investigations. Annotations are updated on a 4-month cycle and new sequence data are imported ~2 months after the release of at least 1000 sequences from any lepidopteran species. Communication with the database curators regarding an imminent release will shorten this time. Metadata linked to each mRNA or EST sequence (life cycle stage, tissue, sex, etc.) have also been databased. Original sequence accession numbers are also listed on each cluster page and linked to EMBL, and can

be searched for with the 'Jump to' search box on the left hand side of every page.

A SHORT TOUR

For security and efficiency reasons, the user-interface pages allow the user to explore the data with certain predefined queries (but see access statement below). ButterflyBase permits simple text searches against the sequence annotation. The definition lines of similar sequences are searched, with the option to define a cut-off value for the precomputed BLAST similarity searches. KEGG (23), GO (24) and EC codes and definitions can also be searched. All searches can be limited to a specific organism or cDNA library.

Once a cluster of interest is found, the cluster page shows a range descriptive data, including the raw data (such as sequence traces if available), the number of ESTs in the cluster, the cDNA libraries they belong to, similarity information from BLAST searches against three databases (Uniref, *Drosophila melanogaster* proteins from FlyBase (25) and *B. mori* predicted proteins from ButterflyBase), and links to the output of all the other BLAST similarity searches. The alignment of the constituent sequences to the consensus can be viewed using an interactive image, a Java applet driven by SEAN or a non-Java text view. These alignment views allow the user to pinpoint databased SNPs. The linked protein page contains basic descriptive data, the predicted sequence, the results of BLAST similarity searches and KEGG, EC, GO and InterPro domain annotation.

EST sequences are a key resource for the development of sequence-specific markers for genetic mapping (26). ButterflyBase facilitates marker development by providing sequence information and a tool for designing degenerate or conserved primers. A protein-driven nucleotide alignment of two orthologous lepidopteran clusters is generated and then used for design of primers using Primer3 (27). EST sequences are also of great utility for the design of microsatellite markers (28). Although transcribed microsatellites are often less polymorphic than non-coding ones (15), they are less likely to be multi-copy or mobile (29). In addition, primers are designed on exon sequences, thus reducing the possibility of null alleles. We provide a simple tool to output any microsatellite present in a specific sequence and also a table of all the microsatellite detected in each species' dataset.

ButterflyBase offers also a BLAST server. Three BLAST search modes are available (NCBI-BLASTALL, PSI-BLAST and WU-BLAST-driven MS-BLAST). MS-BLAST (30) allows a user to query protein databases with multiple short peptide sequences derived from high-throughput mass spectrometry data. PSI-BLAST is particularly effective in the detection of distant similarity and will become an important method for detecting lepidopteran homologues of target genes as the database grows. For more complex queries, a database dump file can be downloaded for local replication of the database, as can species-specific FASTA files of the nucleotide cluster consensus and protein predictions, and custom-built annotation databases used in ButterflyBase.

All datasets, including a SQL flatfile of the database are provided for download with their checksum codes. We also provide FASTA files of some of the custom sequence databases used to carry out similarity searches. One drawback of public EST data, however, is the lack of a raw sequence trace repository. PartiGene can utilize these traces to assist the Phrap alignments, but we are also using them to check manually for the quality of specific libraries or clusters of interest. For this reason, all sequence traces we process are publicly available for download from their respective cluster pages along with a short text file on how the sequence was processed by trace2dbest. This is, unfortunately, only available for sequence trace data we have access to, namely *Heliconius* sp. and *B. anynana*. We are, however, encouraging the community to submit to us their raw sequence data.

SUMMARY OF CONTENT AND UTILITY

Website usage is outlined in the online User's Manual but a summary of the content follows. The main webpage provides an up-to-date overview of the content of the database. At the time of print, ButterflyBase has processed 273 077 mRNA sequences from 32 lepidopteran species belonging to a total of 12 families giving circa 71 000 gene and almost as many protein objects. Although most of the sequences are from *B. mori*, there are nonetheless now 17 species with more than 500 sequences, and 12 species with more than 1000, representing a valuable comparative dataset (Table 1). Nearly half of the ButterflyBase clusters have similarity to known proteins outside the Lepidoptera clade. Although identity of sequence does not necessarily translate into identity of function, sequence similarity is a first step towards gene finding in this taxon. Also, ~58% of the genes in ButterflyBase are significantly similar to at least one more ButterflyBase species, thus facilitating annotation and the design of degenerate or conserved markers. What is also apparent is the relatively high proportion of Lepidoptera-specific genes, about one-third of the clusters have hits only in sequences derived from Lepidoptera but in *B. mori* (which is the most complete dataset) the proportion is about half of the gene objects (Table 1). The number of gene objects is an overestimate of the exact number of actual genes due to the nature of EST datasets and the lack of a genome backbone. Thus, two sets of ESTs from the same gene will appear as two unigenes if they do not overlap, however, accuracy will increase as sequence information from more Lepidoptera is provided. Furthermore, the whole of the *B. anynana* dataset and ca. 16% of the *B. mori* dataset contains 3′ sequences. Therefore, these gene objects may contain long untranslated regions (UTRs) which are not conserved. In any case, these observations warrant an in-depth investigation and any putative Lepidoptera-specific genes need to be examined in a phylogenetic context in order to determine if they have evolved novel functions specific to Lepidoptera or if they have retained ancestral functions despite gross sequence divergence on the protein level.

Table 1. The content of ButterflyBase (September 2007)

Species (ButterflyBase Code)	Taxon/Family	Proteins @ NCBI[a]	mRNAs @ Bbase	Gene objects @ BBase	Similar to known proteins[b]	Only exist in Lepidoptera[c]	Found in 2+ ButterflyBase species[d]	Clusters with putative SNPs (total SNPs)[e]
Total: 33	Lepidoptera	6907	273 077	70 867	37 962	25 204	9583 (41 093)	4821 (27 808)
Anagasta (*Ephestia*) *kuehniella* (AKC)	Pyralidae	3	28	23	14	6	5 (14)	5 (14)
*Antheraea polyphemus** (ALC)	Saturniidae	45	22	17	17	0	0 (17)	N/A
Antheraea mylitta (AMC)	Saturniidae	51	3912	1433	943	509	535 (1432)	47 (140)
*Antheraea pernyi** (APC)	Saturniidae	65	40	37	37	0	0 (37)	N/A
Antheraea yamamai (AYC)	Saturniidae	35	610	325	157	82	88 (226)	9 (19)
Bicyclus anynana (BAC)	Nymphalidae	11	9848	5726	2375	1207	1012 (3099)	81 (234)
Bombyx mori (BMC)	Bombycidae	3623	184 577	35 876	17 162	19 174	4776 (17 194)	3756 (22 445)
Bombyx mandarina (BNC)	Bombycidae	54	261	205	105	97	90 (194)	3 (3)
Choristoneura fumiferana (CFC)	Tortricidae	74	652	618	359	82	72 (379)	N/A
Euclidia glyphica (EGC)	Noctuidae	N/A	570	259	138	2	2 (122)	18 (50)
Galleria mellonella (GMC)	Pyralidae	95	93	84	68	8	4 (65)	N/A
Helicoverpa armigera (HAC)	Noctuidae	207	1221	733	634	53	50 (663)	19 (118)
*Hyalophora cecropia** (HCC)	Saturniidae	57	20	16	16	0	0 (16)	N/A
Heliconius erato (HEC)	Nymphalidae	157	17 573	6859	4787	1118	856 (5019)	464 (3236)
Heliconius melpomene (HMC)	Nymphalidae	443	4976	1965	1262	408	422 (1531)	99 (369)
*Heliothis virescens** (HVC)	Noctuidae	152	90	83	83	0	0 (83)	N/A
*Helicoverpa zea** (HZC)	Noctuidae	80	40	38	38	0	0 (38)	N/A
Lonomia obliqua (LOC)	Saturniidae	133	1635	671	503	60	58 (514)	25 (63)
Manduca sexta (MSC)	Sphingidae	582	3683	2291	1256	412	301 (1469)	22 (56)
Ostrinia nubilalis (ONC)	Crambidae	146	1761	543	309	137	133 (418)	40 (162)
*Pieris brassicae** (PBC)	Pieridae	17	5	5	5	0	0 (4)	N/A
Papilio dardanus (PDC)	Papilionidae	14	708	307	236	22	20 (248)	27 (102)
Plodia interpunctella (PIC)	Pyralidae	47	6219	3788	1879	483	414 (2079)	28 (80)
Papilio xuthus (PUC)	Papilionidae	41	25	24	24	0	0 (24)	N/A
Plutella xylostella (PXC)	Plutellidae	188	1286	1021	701	108	124 (747)	3 (11)
Samia cynthia spp.* (SCC)	Saturniidae	49	27	27	27	0	0 (27)	N/A
*Spodoptera exigua** (SEC)	Noctuidae	64	48	42	42	0	0 (42)	N/A
Spodoptera frugiperda (SFC)	Noctuidae	241	31 538	6993	4172	1116	1204 (4741)	149 (528)
Spodoptera litura (SLC)	Noctuidae	66	154	100	85	7	8 (90)	1 (1)
*Spodoptera littoralis** (STC)	Noctuidae	28	23	20	20	0	0 (20)	N/A
Tineola bisselliella (TBC)	Tineidae	1	921	240	170	39	14 (162)	30 (177)
Trichoplusia ni (TNC)	Noctuidae	138	511	498	338	74	61 (379)	N/A

*designates those species with no public ESTs but public full-length mRNA sequences.
[a]Nuclear sequences only, this total includes segmented sequences and is not limited to RefSeq. August 2007. The *B. mori* proteins were limited to 1025 before January 2007.
[b]BLASTx of nucleotide consensus and BLASTp of predicted proteins versus Uniref100 or proteins released by the *Apis mellifera*, *D. melanogaster*, *Tribolium castaneum* and *Anopheles gambiae* genomes or other Arthropoda proteins in EBI with *E*-value cutoff $1E-4$ (source: EBI Jul 2007). We also used in-house clusters of the public EST data for *Aedes aegypti, Anopheles gambiae, Culex pipiens, Drosophila ananassae, Drosophila erecta, Drosophila grimshawi, Drosophila simulans, Drosophila yakuba* and *Tribolium castaneum* (*E*-value cutoff $1E-4$, source: EBI September 2007).
[c]BLASTn of nucleotide consensus versus Lepidoptera nuclear nucleotides, *B. mori* genome from EBI and ButterflyBase EST consensuses but no significant similarity to the databases mentioned above (EBI, Jul 2007, *E*-value cutoffs $1E-4$).
[d]Lepidoptera-specific clusters which were found to have a significant hit in at least one other organism in ButterflyBase using BLASTn for nucleotide consensuses or BLASTp for protein predictions (Jul 2007, *E*-value cutoff $1E-3$). Gene objects present in more than one organism facilitate annotation and marker design. In brackets, a similar count is present for all clusters regardless of similarity to any protein.
[e]Most Lepidoptera cDNA libraries are constructed with relative outbred individuals, thus the relatively high number of SNPs. Even though the number of clusters containing putative SNPs are accurate, the reader has to consider that the total number of SNPs may be inflated as the data here are pooled from all cDNA libraries.

Phylogenetics

The phylogenetic context of Lepidoptera is one of the taxon's strongest advantages for the study of ecology and evolution. Although the amount of public genomic data in Lepidoptera is increasing rapidly, the phylogenetic coverage is limited to the Ditrysia and non-existent for basal clades. A broader phylogenetic sampling, of at least a handful of chosen genes will help improve much of the unresolved lepidopteran phylogeny and also shed more light on the evolutionary dynamics of Lepidoptera-specific genes. Different levels of phylogenetic investigation require different kinds of genes, thus fast-evolving genes are only suited for building phylogenies of closely related species whereas highly conserved genes (such as ribosomal proteins) are best suited for inferring the relationships among the more basal lineages. A broad phylogenetic analysis of ~300 species using up to 26 genes derived from EST sequences is already underway (Leptree.net; Mitter, personal communication) and the tools developed in ButterflyBase will facilitate this and similar research.

Annotation

ButterflyBase is primarily an annotation platform. Currently, the only information provided is similarity to

known sequences, including to other lepidopteran sequences. The aim of the annotation platform is to host enough information to allow researchers to judge if their sequence of interest has a specific annotation identity. This annotation will be essential for annotating novel sequences especially short reads generated in some projects such as cDNA–AFLPs. Currently, we do not provide curated annotation information but in the near future we will publish analysis on orthologue groupings. We plan to allow the community itself to contribute annotations for each ButterflyBase object perhaps by using a Wiki-based annotation platform (31) or the Generic Model Organism Database toolkit (GMOD). In addition, we hope to expand the annotation platform to include both non-EST sequence data and genetic/phenotypic data within 2008. Such an effort will be initialized by a conversion to the more standardized database schema of Chado from the GMOD (32). The major obstacle is however the lack of a fully sequenced genome with which to anchor the genomic data. The quality of the first releases of *B. mori* is not sufficient for the purpose but a joint assembly is expected to be made public within 2007. With a GMOD-compatible database and a *B. mori* genome the capability of ButterflyBase as an annotation tool will be greatly enhanced. Likewise, as additional EST datasets are made public, the quality of the annotation will increase.

Linkage mapping and molecular evolution

ButterflyBase was originally developed for the generation of EST-based molecular markers for *Heliconius* sp. (26,33). Using ButterflyBase data, a researcher may generate conserved, degenerate or species-specific markers of specific single-copy genes. Pringle *et al.* (33) used this approach to provide the first extensive evidence for conserved macro-synteny between *H. melpomene* (a butterfly) and *B. mori* (a moth), two species whose sequence divergence has reached saturation in third codon positions. ButterflyBase provides also predicted SNPs, which have been determined from the clustered alignment. These identified SNPs (and RFLPs) can be verified by visual inspection of the alignment. Such data allow the generation of SNP-based markers to survey natural populations for association mapping projects or estimate the rate of evolution of specific proteins. Researchers using a cDNA approach to acquire SNP information for linkage mapping can also make use of ButterflyBase's services and in the process contribute to the pool of public sequence information for Lepidoptera.

Proteomics

An important function of genomic datasets is to guide future biochemical investigations. In taxa such as Lepidoptera, where much of the proteome is unknown and composed of many previously unidentified genes, *de novo* protein sequencing provides valuable information. In such proteomes, standard methodologies for identifying peptides by mass spectrometric (MS) data are more error-prone and can be misleading. The MS-BLAST server facilitates identification using the ButterflyBase predicted (and often partial) proteins.

Support small-scale sequencing

During the construction of ButterflyBase we used all available Lepidoptera ESTs hosted in the public domain. A fraction of them was unfortunately lacking information, or contained vector contamination and/or low-quality sequence. ButterflyBase provides the facility to host trace information and currently holds raw trace data from *H. erato*, *H. melpomene* and *B. anynana*. In the future, ButterflyBase's pipeline will judge the quality of a cDNA library based on the number of errors as detected from ESTs from other libraries or published full-length mRNAs. This is only possible, however, for species where multiple libraries of sufficient depth exist. In addition, ButterflyBase can offer the service of processing raw traces and generate dbest submission reports to researchers who request so and thus allow for a more standardized collection of Lepidoptera sequence information. In the near future, a new international Advisory Board will guide ButterflyBase and will post a set of recommendations for submissions of data to GenBank.

DATA SUBMISSION AND ACCESS STATEMENT

All ButterflyBase data are freely and publicly accessible. To be included in ButterflyBase, EST and mRNA data should be submitted to EMBL/GenBank/DDBJ (a step which we can handle upon request). We strongly encourage submission of raw trace files (in SCF format) to ButterflyBase. Although the user is limited to predefined queries and can download a copy of the database, we can also run custom queries upon request (email query at butterflybase.org). Our goal for the future is to develop the project guided by the community. Therefore, we welcome requests and contributions.

ACKNOWLEDGEMENTS

We are grateful to the Blaxter Neglected Genomes bioinformatics team for support and use of compute resources, especially Ann Hedley and Ralf Schmid. Hendrik Tilger, Martin Niebergall and Dieter Ruder set up the distributed computing. Walter Traut, Mathieu Joron, Simon Baxter, Jim Mallet, Patricia Beldade and David G. Heckel provided many useful comments. Mathieu Joron created the first EST dataset with which ButterflyBase used for its development. A.P. and S.G.J. are supported by the Max Planck Gesellschaft (Germany), W.O.M. by the American National Science Foundation and NEScent, the National Evolutionary Synthesis Center, C.D.J. by a Royal Society Fellowship (UK) and M.L.B. by the Natural Environment Research Council (NERC, UK). Initial support was provided by the Biotechnology and Biological Sciences Research Council (BBSRC, UK). Author contributions: The initial *Heliconius* EST database was conceived by C.J. and M.B. The extension from the '*Heliconius* ButterflyBase' to 'ButterflyBase' was conceived and developed by A.P. with

additional technical support from S.G.J. Intellectual support and motivation was from W.O.M. This article was drafted by A.P., M.L.B., W.O.M. and C.J. All authors approved the final version of the manuscript. Funding to pay the Open Access publication charges for this article was provided by Max Planck Gesellschaft (Germany).

Conflict of interest statement. None declared.

REFERENCES

1. Xia,Q., Zhou,Z., Lu,C., Cheng,D., Dai,F., Li,B., Zhao,P., Zha,X., Cheng,T. *et al.* (2004) A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science*, **306**, 1937–1940.
2. Mita,K., Kasahara,M., Sasaki,S., Nagayasu,Y., Yamada,T., Kanamori,H., Namiki,N., Kitagawa,M., Yamashita,H. *et al.* (2004) The genome sequence of silkworm, *Bombyx mori*. *DNA Res.*, **11**, 27–35.
3. Wu,C., Asakawa,S., Shimizu,N., Kawasaki,S. and Yasukochi,Y. (1999) Construction and characterization of bacterial artificial chromosome libraries from the silkworm, *Bombyx mori*. *Mol. Gen. Genet.*, **261**, 698–706.
4. Yasukochi,Y., Ashakumary,L.A., Baba,K., Yoshido,A. and Sahara,K. (2006) A second-generation integrated map of the silkworm reveals synteny and conserved gene order between lepidopteran insects. *Genetics*, **173**, 1319–1328.
5. Yamamoto,K., Narukawa,J., Kadono-Okuda,K., Nohata,J., Sasanuma,M., Suetsugu,Y., Banno,Y., Fujii,H., Goldsmith,M.R. *et al.* (2006) Construction of a single nucleotide polymorphism linkage map for the silkworm, *Bombyx mori*, based on bacterial artificial chromosome end sequences. *Genetics*, **173**, 151–161.
6. Miao,X.X., Xub,S.J., Li,M.H., Li,M.W., Huang,J.H., Dai,F.Y., Marino,S.W., Mills,D.R., Zeng,P. *et al.* (2005) Simple sequence repeat-based consensus linkage map of *Bombyx mori*. *Proc. Natl Acad. Sci. USA*, **102**, 16303–16308.
7. Yoshido,A., Bando,H., Yasukochi,Y. and Sahara,K. (2005) The *Bombyx mori* karyotype and the assignment of linkage groups. *Genetics*, **170**, 675–685.
8. Beldade,P., McMillan,W.O. and Papanicolaou,A. (2007) Butterfly genomics eclosing. *Heredity* [Epub ahead of print].
9. Bouck,A. and Vision,T. (2007) The molecular ecologist's guide to expressed sequence tags. *Mol. Ecol.*, **16**, 907–924.
10. Beldade,P., Rudd,S., Gruber,J.D. and Long,A.D. (2006) A wing expressed sequence tag resource for *Bicyclus anynana* butterflies, an evo-devo model. *BMC Genomics*, **7**, 130.
11. Cheng,T.C., Xia,Q.Y., Qian,J.F., Liu,C., Lin,Y., Zha,X.F. and Xiang,Z.H. (2004) Mining single nucleotide polymorphisms from EST data of silkworm, *Bombyx mori*, inbred strain Dazao. *Insect Biochem. Mol. Biol.*, **34**, 523–530.
12. Mita,K., Morimyo,M., Okano,K., Koike,Y., Nohata,J., Kawasaki,H., Kadono-Okuda,K., Yamamoto,K., Suzuki,M.G. *et al.* (2003) The construction of an EST database for *Bombyx mori* and its application. *Proc. Natl Acad. Sci. USA*, **100**, 14121–14126.
13. Negre,V., Hotelier,T., Volkoff,A.N., Gimenez,S., Cousserans,F., Mita,K., Sabau,X., Rocher,J., Lopez-Ferber,M. *et al.* (2006) SPODOBASE: an EST database for the lepidopteran crop pest *Spodoptera*. *BMC Bioinformatics*, **7**, 322.
14. Parkinson,J., Anthony,A., Wasmuth,J., Schmid,R., Hedley,A. and Blaxter,M. (2004) PartiGene—constructing partial genomes. *Bioinformatics*, **20**, 1398–1404.
15. Thiel,T., Michalek,W., Varshney,R. and Graner,A. (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare L.*). *Theor. Appl. Genet.*, **106**, 411–422.
16. Huntley,D., Baldo,A., Johri,S. and Sergot,M. (2006) SEAN: SNP prediction and display program utilizing EST sequence clusters. *Bioinformatics*, **22**, 495–496.
17. Parkinson,J., Guiliano,D.B. and Blaxter,M. (2002) Making sense of EST sequences by CLOBBing them. *BMC Bioinformatics*, **3**, 31.
18. Ewing,B., Hillier,L., Wendl,M. and Green,P. (1998) Basecalling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, **8**, 175–185.
19. Wasmuth,J.D. and Blaxter,M.L. (2004) prot4EST: translating expressed sequence tags from neglected genomes. *BMC Bioinformatics*, **5**, 187.
20. Iseli,C., Jongeneel,C.V. and Bucher,P. (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 138–148, http://www.ch.embnet.org/software/ESTScan.html.
21. Fukunishi,Y. and Hayashizaki,Y. (2001) Amino acid translation program for full-length cDNA sequences with frameshift errors. *Physiol. Genomics*, **5**, 81–87.
22. Zdobnov,E.M. and Apweiler,R. (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.
23. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.
24. Gene Ontology Consortium. (2006) The gene ontology (GO) project in 2006. *Nucleic Acids Res.*, 34, D322–D326.
25. Crosby,M.A., Goodman,J.L., Strelets,V.B., Zhang,P. and Gelbart,W.M. (2007) FlyBase: genomes by the dozen. *Nucleic Acids Res.*, **35**, D486–D491.
26. Papanicolaou,A., Joron,M., McMillan,W.O., Blaxter,M.L. and Jiggins,C.D. (2005) Genomic tools and cDNA derived markers for butterflies. *Mol Ecol.*, **14**, 2883–2897.
27. Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, **132**, 365–386.
28. Woodhead,M., Russell,J., Squirrell,J., Hollingsworth,P.M., Mackenzie,K., Gibby,M. and Powell,W. (2005) Comparative analysis of population genetic structure in *Athyrium distentifolium* (Pteridophyta) using AFLPs and SSRs from anonymous and transcribed gene regions. *Mol. Ecol.*, **14**, 1681–1695.
29. Zhang,D.-X. (2004) Lepidopteran microsatellite DNA: redundant but promising. *Trends Ecol. Evol.*, **19**, 507–509.
30. Shevchenko,A., Sunyaev,S., Loboda,A., Shevchenko,A., Bork,P., Ens,W. and Standing,K.G. (2001) Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Anal. Chem.*, **73**, 1917–1926.
31. Salzberg,S.L. (2007) Genome re-annotation: a wiki solution? *Genome Biol.*, **8**, 102.
32. Mungall,C.J. and Emmert,D.B. FlyBase Consortium (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, **23**, i337–i346.
33. Pringle,E.G., Baxter,S.W., Webster,C.L., Papanicolaou,A., Lee,S.F. and Jiggins,C.D. (2007) Synteny and chromosome evolution in the Lepidoptera: evidence from mapping in *Heliconius melpomene*. *Genetics*, **177**, 417–426.

Chapter 4 - The GMOD Drupal Bioinformatic Server Framework

This Chapter produced the first bioinformatic library within the Drupal Content Management System (CMS). Included was i) a library for manipulating Chado and GMOD data (gmod-dbsf), ii) an innovative annotation server (biosoftware_bench) and iii) a module to database and disseminate RNAi experiments (genes4all_experiment). All are deployed within InsectaCentral and the latter was used in a recent review by Terenius et al 2010.

Citation

Papanicolaou, A. Heckel, D.H. Accepted in Bioinformatics (Oxford) on 23rd of September 2010.

The GMOD Drupal Bioinformatic Server Framework

Papanicolaou A. [1 2 3*] and David G. Heckel [2]

[1] Centre for Conservation and Ecology, University of Exeter in Cornwall, Penryn TR10 9EZ, United Kingdom

[2] Department of Entomology, Max Planck Institute for Chemical Ecology, Hans-Knöll Str 8, Jena D-07745, Germany

[3] CSIRO Ecosystem Sciences, Black Mountain Laboratories, Clunies Ross St, Acton 2601, Australia

Associate Editor: Prof. Alfonso Valencia

ABSTRACT

Motivation: Next Generation Sequencing technologies have led to the widespread use of -omic applications. As a result, there is now a pronounced bioinformatic bottleneck. The General Model Organism Database (GMOD) tool kit (http://gmod.org) has produced a number of resources aimed at addressing this issue. It lacks, however, a robust online solution that can deploy heterogeneous data and software within a Web Content Management System (CMS).

Results: We present a bioinformatic framework for the Drupal CMS. It consists of 3 modules. First, GMOD-DBSF is an Application Programming Interface module for the Drupal CMS that simplifies the programming of bioinformatic Drupal modules. Second, the Drupal Bioinformatic Software Bench (biosoftware_bench) allows for a rapid and secure deployment of bioinformatic software. An innovative graphical user interface (GUI) guides both use and administration of the software, including the secure provision of pre-publication datasets. Third, we present genes4all_experiment, which exemplifies how our work supports the wider research community.

Conclusion: Given the infrastructure presented here, the Drupal CMS may become a powerful new tool set for bioinformaticians. The GMOD-DBSF base module is an expandable community resource that decreases development time of Drupal modules for bioinformatics. The biosoftware_bench module can already enhance biologists' ability to mine their own data. The genes4all_experiment module has already been responsible for archiving of more than 150 studies of RNAi from Lepidoptera, which were previously unpublished.

Availability and Implementation: Implemented in PHP and Perl. Freely available under the GNU Public License 2 or later from http://gmod-dbsf.googlecode.com

Contact: alexie@butterflybase.org

1 INTRODUCTION

1.1 Emerging model species and bioinformatics

Next generation sequencing (NGS) technologies have allowed an increasing number of biologists to utilize the -omic experimental strategy and support research programs by searching for statistically significant patterns in Large Scale (LS) experiments (Collins et al., 2003). Due to the limited number of bioinformaticians and resources, this rapid uptake of -omics is now causing a bioinformatic bottleneck. This bottleneck, which is more pronounced in the Ecological and Evolutionary Functional Genomics (EEFG) community (Beldade et al., 2008), ought to be addressed without requiring custom-made and non-integrated solutions. The Generic Model Organism Database tool-kit (GMOD; http://gmod.org) is a consortium originally formed from functional genomics model organism communities to produce a standard set of open-source software for handling, primarily, genomic data. Since its inception, the consortium has built or incorporated an impressive array of tools and standards. The uptake of GMOD tools and standards has been so successful that GMOD has expanded beyond the functional genomics community and is now been used by EEFG laboratories.

Indeed, 'MOD' databases are now commonplace in the -omics field (http://gmod.org/wiki/GMOD_Users). Until recently, GMOD software had focused on whole genome sequencing. As researchers from other fields make use of -omic, bioinformatic and Artificial Intelligence (AI) approaches, GMOD has expanded into other fields such as phylogenetics (Heinicke et al., 2007), microarray research (Day et al., 2007), molecular ecology (a Chado extension from the National Synthesis Center for Evolution; http://www.nescent.org/informatics/software.php), transcriptomics without a reference genome (Papanicolaou et al., 2009) and others. Further, the cost-effectiveness of NGS does not apply to the downstream cost associated with computational analysis of the data; quite the opposite, in fact. Therefore, there is an ever-growing need for cost-effective and integrated solutions that improve the capabilities of wet-lab biologists to mine their own data before publication. Even though a number of commercial tools exists, some are not affordable. Others are closed-source software and thus cannot be adapted. Moreover, most are not integrated into the larger GMOD framework. Individual attempts within the GMOD consortium have yet to provide a generic visualization front end. The website creation tool, GMODWeb (O'Connor et al., 2008) is of interest but it has limited scope but it is useful in rapidly

*To whom correspondence should be addressed.

generating a web-based front-end for a Chado database (Mungall et al., 2007; Arnaiz et al., 2006). Tripal (http://gmod.org/Tripal) offers an efficient front-end for Chado but no generic framework. InterMine (Lyne et al., 2007) is a more powerful Graphical User Interface (GUI) for a database, driven by lightweight JavaScript but it is a complicated framework to use for development. The Ensembl system (Hubbard et al., 2002) is an example of a complete platform for processing genomic data but it was custom built for the needs of the Sanger Institute rather than a community software. Indeed, most of above software are open-source but not necessarely developed for open-development. In order to minimize reliance on continued funding, the community could orientate towards more generic frameworks explicitly designed for open-development. Bioinformatic work-flow visualizations, such as Taverna and Galaxy (Oinn et al., 2004; Giardine, 2005), are both geared towards data analysis, even though the latter allows for custom plugins. Although the Galaxy team is working towards a more general framework for data dissemination, for many bioinfomaticians, the Ensembl solution seems more robust. Ensembl is an entire bioinformatic framework with both analysis and dissemination tools, but it has a very steep learning curve and is not a GMOD component. It would be of interest, however, for the entire GMOD community to develop a generic 'plumbing' framework so that i) laboratories can rapidly deploy websites with data analysis/dissemination tools (such as Taverna or a BLAST server); ii) bioinformaticians can rapidly program new applications (such as custom front-ends on par with InterMine).

1.2 The Drupal Content Management System

One solution is to use a Content Management System (CMS) such as Wordpress, Drupal or others. CMSs are platforms for storing, managing, disseminating data of any type. Often they have been used to drive websites, including 'blogs', but research projects such as Scratchpads (Smith et al., 2009) have also been successful. Some researchers use CMSs for building their laboratory websites. Some CMSs support a number of useful concepts such as the RDF, XML and similar protocols, ontologies, controlled vocabularies and the other concepts relating to the Semantic Web. Further, CMSs are often a complete software package with tools for managing community-based data, such as users, roles and fine-grained permissions. Some CMSs are modular, allowing for users to program their own plugins and extend functionalities. Drupal is such a CMS. It is open-source and can be downloaded freely from http://drupal.org. It is written in PHP, a language that is straightforward for nascent bioinformaticians to learn. It supports a number of database engines, including mySQL, Oracle and the GMOD supported PostgreSQL. Further, Drupal is built with security in mind, has powerful user-management tools and is highly modular, allowing for plugins to be developed and deployed in a standardized and streamlined fashion. Importantly, Drupal is popular and well documented. The widespread use has resulted in a large active community of users and developers (e.g. see http://egressive.com/article/who-uses-drupal).

This paper initiates a long-term effort in creating a bioinformatic framework for the Drupal CMS within the specifications of GMOD. We developed three modules for three categories of users. First, GMOD-DBSF is a generic function framework for Drupal developers of bioinformatic tools. Then we built two modules for end-users: i) a powerful similarity-search software (e.g. BLAST) server for wet-lab biologists and system administrators benefiting from a friendly GUI and ii) an RNAi experiment databasing platform. The latter can be easily modified for other experimental data, but it was developed and used in a community-wide review on failed and unpublished RNAi experiments in Lepidoptera (butterflies and moths; the only taxon where RNAi experiments are often unsuccessful).

2 METHODS

We used the Drupal 6 CMS. As the Chado package uses PostgreSQL, this database engine is required. The GMOD-DBSF module is a base-module and thus required for all other modules in this framework; the other modules are optional. The GMOD-DBSF base-module can utilize an installation of the Chado package but installing it is not necessary as it is not required for the biosoftware_bench module. The BioPerl (http://bioperl.org; Stajich et al., 2002) package and freely available Perl libraries (from CPAN) are needed along with certain Drupal modules: the Tabs module (http://drupal.org/project/tabs) is used to deploy tabular web content; the JQuery module (http://drupal.org/project/jquery) to deploy and seamlessly maintain the JQuery JavaScript library. Further, an external JQuery-utilizing library, dynatree (http://code.google.com/p/dynatree), is used to produce "check-box trees". Commonly used annotation software, such as BLAST, annot8r, InterProScan and SSAHA2 (Altschul et al., 1997; Schmid and Blaxter, 2008; Zdobnov and Apweiler, 2001; Ning et al., 2001), were integrated into biosoftware_bench, but using them requires that they are installed on the server (not all software needs to be installed: administrators can select which ones they wish to make available). For sequence retrieval, the fastacmd and Bio::Seq::Index approaches were used for the BLAST and SSAHA2 databases, respectively. To enable job-management, Condor (Thain et al., 2005) was used as it is simple and can perform well on both a PC-farm and a single multi-core host. Future versions of this framework aim to make use of the Sun Grid Engine.

2.1 Specifications

The framework adheres to certain criteria. It i) is open-source under a non-restrictive license and thus can be customized and expanded; ii) can be integrated with other widely used bioinformatic applications and implement the GMOD standards; iii) is secure to both the user and the server; iv) provides GUIs to both end-users and administrators; v) is developer friendly by extending Drupal's Application Programming Interface (API) according to the Drupal community specification. Drupal itself has a powerful API: e.g. the deployment of the 3rd party modules, such as the ones presented here, requires no more than a single line of code. This complements their installation, which is usually "point-and-click".

2.2 Aims

The work presented here focused on producing three Drupal modules. The first is GMOD-DBSF, which provides a framework for developing new bioinformatic Drupal modules. It is responsible for i) importing a subset of the Chado tables to Drupal, ii) creates new tables in Drupal using Chado conventions; iii) provides functions to communicate with Chado and Drupal database schemas; iv) provides other, generic, functions useful for bioinformatic module development.

We also built two example applications. First, a software server with the BLAST, InterProScan, annot8r, and SSAHA2 software deployed by default; additional plugins can be generated by the community. Second, a web-based database for storing experimental information from RNAi experiments. We used the Minimum Information Criteria for RNAi experiments (MIARE) as provided by the MIARE working group (http://miare.sourceforge.net) and the Lepidoptera RNAi Working Group, an international group composed of 70 scientists from 42 institutions in 21 countries (Terenius et al., in press).

2.3 Schema

We opted not to use Chado for public data entry and manipulation; in our work, Chado is used as a long-term and secure data warehouse. We prefer not to allow the public to commit changes to the Chado database but still wish to provide a bidirectional user-interface. We, therefore, use Chado for read operations of data residing in the data warehouse but opted to create a Chado schema within Drupal for read/write operations of user-contributed data. With Chado being a highly generic schema, there are a number of tables unused in this instance of GMOD-DBSF (e.g. the MAGE module). Therefore, we imported, therefore, only the basic Chado tables in the Drupal database (the feature, organism, cv, dbxref, pub tables and their dependencies). Drupal is then extended with additional tables created using the Chado conventions (Figure 1). New tables in the 'resource' group were created to allow better representation of sequence-less features. Likewise, a software table group is utilized specifically for software variables and is linked with the resource data using the software_resource. Further, a new study group of tables has been created to allow for generalized databasing of wet-lab experimental data. Publications are supported via the Chado pub schema. We used new tables to better integrate authorships using the author and pub_author tables. This implementation allows seamless integration with core Drupal data: e.g. a resource_roles allows linking of the resources with specific Drupal username groupings (roles). All of these tables are installed automatically during the point-and-click installation of GMOD-DBSF. It was expected that certain applications would require the synchronization of data between the Drupal and Chado databases. For example, InsectaCentral requires it for its Community Annotation module. For the security of Chado as a data-warehouse, developers should be cautious but secure protocols can be developed using Drupal's features. GMOD-DBSF offers such feature/resource-specific synchronization. In InsectaCentral's implementation, a special administrator user-group is allowed to use these functions and synchronization changes with Chado.

2.4 RNAi experiment

In order to efficiently provide a cataloguing platform for the RNAi experiments, the Lepidoptera RNAi working group used the MIARE ontologies. MIARE is a set of reporting guidelines that describes the minimum information that should be reported about an RNAi experiment to enable the unambiguous interpretation and reproduction of the results. We then built the genes4all_experiment module using GMOD-DBSF and enhanced it via community feedback. Three different data-types are used: sequence features, sequence-less features and publications. To distinguish between the first two, the latter is called a resource and has a separate set of tables in our schema. Two types of sequence data are used: i) the target gene, which may be derived from a species other the one targeted (due to lack of sequence information), and ii) the RNAi construct. Three types of resources are used: i) experimental animals, ii) delivery protocol and iii) assay protocol. Considering that the genes4all_experiment caters primarily to unpublished research, the publication GUI requests only the communicating author but, as we mentioned above, the schema can handle multiple authors and their details via the author and pub_author tables.

3 RESULTS

3.1 Drupal for bioinformatics using GMOD-DBSF

The core Drupal program has limited capabilities for bioinformatics. As a CMS, it is most capable in storing, displaying and organizing data as stored in the so-called “nodes”: authored web-pages linked with ancillary data. Extensions, called “modules”, extend its functionality. For example, the Tabs module that we use allows for multiple web-pages to appear as tabs. Such modules provide their own API and thus allow other modules to make use of a complicated functionality using only a line of code. The GMOD-DBSF module is one such module. Bioinformaticians can use it to perform an increasing number of operations (see http://gmod-dbsf.googlecode.com/files/GMOD-DBSF_dev_manual_1.0.pdf). Indeed, we hope that as the bioinformatics community embraces Drupal, GMOD-DBSF will also expand.

Currently, GMOD-DBSF offers a number of functionalities not available in the Drupal core. A set of functions allows a generic interaction with Chado tables. The function *gmod_dbsf_add_cv()*, for example, allows for one to add a new Controlled Vocabulary (CV) by providing the name of the CV and an array with the CV terms they wish to add. This function can connect to a Chado database via the *gmod_dbsf_db_execute()* function or operate on the local Drupal database (or make use of the *gmod_dbsf_is_chado()* auto-detect function). Similar functions operate to add, delete and populate the feature, db, pub and other Chado tables. Ancillary Chado tables, such as the featureprop and feature_cvterm tables, often require complicated SQL commands with multiple joins. A number of gmod_dbsf functions cater to simplify manipulating these tables by simply passing the desired variables. For example a featureprop table can be populated with a single line of code which passes the feature ID or feature name, the CV term and properties one wishes to associate. This approach is the *raison d'etre* of GMOD-DBSF: to allow other modules to query and manipulate Chado in a standardized fashion, and also to accelerates the development of other modules. Other convenience functions allow a developer to install a materialized view, a new table or PostgreSQL function. A few functions aim to provide secure approaches for oft-used tasks. The *gmod_dbsf_create_uid()* function (all non-core functions in Drupal begin with the module's name) creates a unique MD5 identifier, based on a user's session ID, time and optionally a text string, which can be used for file uploads. The *gmod_dbsf_batch_upload_fasta()* function allows users to upload a FASTA file to the server even if it is many megabytes or takes a considerable amount of time. It is used, for example, by the biosoftware_bench software server to allow users to upload datasets for use as query or subject databases. Finally, a few functions have been created to make use of BioPerl functions. For example, one function is responsible for creating and parsing GFF3 files, another, the *gmod_dbsf_get_taxonomy_from_ncbi()*, uses Bio::DB::Taxon to query NCBI (via Entrez or via a local NCBI Taxonomy database flatfile) for the taxonomy of a species. In InsectaCentral, this function is used in conjunction with the *gmod_dbsf_get_add_organism()* function to build a GUI for InsectaCentral curators to add new organisms and ancillary phylogenetic information into the Chado database.

3.2 Bioinformatic Software Bench

Innovations

When a laboratory generates multiple pre-publication datasets, a local solution for mining, searching and manipulating the data must be deployed. This leads to cumbersome administration and maintenance and the need for constant bioinformatic support. There are a number of main innovations of biosoftware_bench: i) graphical administration; ii) deployment of command line software; iii) use of a secure daemon to handle job submissions with the option to use the Condor job management system, and iv) linking datasets with phylogenetic information. Further, the ability

to deploy datasets only available to certain users or groups allows for the existence of a single server to handle both public and pre-publication data. As the system is integrated with a laboratory's website and user authorization is handled by Drupal, the entire process appears seamless to the user. Moreover, the deployed software can also be used by other modules, i.e. without a GUI. By re-using the same biosoftware_bench functions, another module can utilize them to prepare and process software results. For example, a module currently under development allows for community members to submit an Open Reading Frame, which is then automatically processed and annotated with the BLAST, annot8r and InterProScan software, with the resulting data stored first into Drupal and then transferred into Chado by a curator.

Installation of software plugins

The module comes with plugins for BLAST, annot8r, InterProScan and SSAHA2 but others can be coded by the community. Bioinformatic software can be installed through the biosoftware_bench module. Two "include" (.inc) files are needed for each software. One file guides the installation, including the use of CV terms to define options. The second file is responsible for the interface and batch jobs. New software can be deployed within a few hours by creating two such files and providing Perl routines to handle any output graphs. Once deployed, administrators have access to a set of options that allows them to select which software they wish to install and if they wish to make use of the Condor job management system. This latter feature allows administrators to utilize a PC-farm or a multi-core server to control job submissions. In both cases, a Perl daemon containing the aforementioned Perl routines, processes the jobs as an unprivileged user. For the software servers, it also post-processes the output of the software search in order to provide the output as a number of file formats. A Bio::Graphics-driven image of an alignment of the hits to the query is also produced and colored according to the significance statistic.

Administration

The biosoftware_bench module provides an administrator's GUI to minimize user-errors and reduce the time required to setup and maintain a software server. In Drupal, administrative rights are decoupled for each module and each action. Users with specific administrative rights have a GUI where they can specify the location of datasets, see which ones are available and choose which to deploy. The administrator can provide friendly names and group memberships (e.g. "Genomes", "Transcriptomes", "UniProt" etc.) to assist users selecting an appropriate dataset. System and security checks prevent errors with typing or dataset-formatting and thus ensure that the database is populated only with functional datasets. Further, linking them to a species through the NCBI Taxonomy database can be used to enable the phylogeny-driven dataset selection. One security feature allows the administrator to decide if a dataset is to be made restricted to a specific set of users. This allows for the secure deployment of both public and pre-publication datasets from the same server and interface. In the future, for large websites biosoftware_bench ought to load the dataset in the database rather than use flatfiles. The current method of providing formatted flatfiles is, however, the most straightforward approach and will suit the bulk of biosoftware_bench administrators, in particular bioinformaticists with limited programming or databasing skills.

End-user capabilities

The privacy mechanism allows end-users to see only the datasets that their username and role memberships allow. In the BLAST server, they can choose to run multiple BLAST algorithms simultaneously, expand their subject dataset by uploading a multi-FASTA file and use a phylogeny to select species- or taxon-specific subject datasets. Once the search is submitted, a self-refreshing page with a unique submission identifier (SUID) appears and can be used to bookmark the page. The system uses 'cron' jobs to purge old files, and administrators can decide when result files are flagged as being old and ready to be deleted. For the BLAST software, the results are first produced as XML but BioPerl modules provide an additional choice of text and HTML output. For other software, when possible, an XML is also provided as well as GFF and/or HTML and text. A Bio::Graphics-driven alignment graph provides an overview of the queries, any hits and their respective scores. The tabular presentation of significant hits allows users to download hits of interest as a FASTA file.

3.3 Experiment module

The experiment module was custom-built for the International Lepidoptera RNAi Working Group (Terenius et al., under review). It utilizes functions provided by GMOD-DBSF. A number of core functions exist and adapting them for other types of experiments is straightforward. The GUI was built to provide a good balance between user-friendliness and data security. Each fully completed submission is live in real-time, partial submissions can be continued later and even completed submissions may be edited by authorized individuals. The date and time of the last submissions/edits is stored in the database. The user is requested to first provide a unique name for their submission, their email address (which is not made public) and a non-unique passkey. The passkey can be used by multiple submissions and its purpose is to prevent unauthorized edits. Further, data linked with a passkey can be reused by subsequent submissions and allows for continuing incomplete ones. After providing these credentials, the user is presented with a panel of 6 tabs. The first tab relates to the sequence-based features: the target gene and RNAi construct. The second tab handles non-sequence data (resources) such as experimental animals and protocols. The third tab contains the publication data, including external database cross-references. Such references are also available for the resources and features, allowing users to identify experimental animal stocks or GenBank gene identifiers. Once all required information is provided, a finalize tabs becomes available and users are able to review their submission prior storing the study as 'complete'. At any time, users can stop and continue their submission at a later day by making use of their passkey credentials.

In order to reduce work load to curators, much of the data is driven by controlled vocabularies (as provided by the community). Building new modules might be considered time-consuming. It might be of interest, therefore, to note that the design and deployment of this module required two weeks of full-time equivalent work, excluding a week of responding to community

feedback. Using the existing module as a template, however, other types of experiments can be supported in a matter of hours.

4 DISCUSSION

4.1 Utility as a community resource

Unlike other software, the work presented here aims to integrate well with an existing laboratory website. This system allows laboratories to deploy software locally. This is especially useful for software that can take advantage of clusters of computers (e.g. the RaxML phylogenetics or PAML molecular evolution software; (Stamatakis et al., 2008; Yang, 2007)). Further, by utilizing a CMS, laboratories can deploy the biosoftware_bench module via the point-and-click approach. They can, likewise, create their entire web content including feed aggregators (e.g. Atom), blogs and file servers. Indeed, a user-friendly system can be the key to allow a specific -omics community (such one centered around a taxon or a genome sequencing project) to develop and interact with a central resource such as a large database supporting that community. Drupal modules offer a straightforward installation but also allow for customization within a variety of existing "themes". It is possible, then, to provide the feeling that the -omic data, BLAST servers and standard web-pages are part of one package.

4.2 Utility as a bioinformatic framework

With the explosion of information and the paucity of expertise, Drupal is already being applied across biological disciplines: recent work funded by the European Union Framework 6 has produced Scratchpads, a Drupal project for Natural History collections (Smith et al., 2009). With advances in information technology and increased interest in semantic integration, the genomics community will benefit from choosing a diverse and robust system, such as Drupal, for integrating, analyzing and displaying information. With more genome sequencing project coming to fruition, there will be laboratories focusing on data-types such as ecological and population data which, thus far, are not part of genome databases. GMOD-DBSF is a step towards addressing these emerging needs without worsening the bioinformatic bottleneck.

This new API for Drupal makes the co-existence of Chado and Drupal seamless to the end-user and reduces the learning curve for the bioinformatic community. Additionally, a large number of core functions or 3rd party modules are available to be used by the bioinformatic community. One example is Drupal's abilities for data federation. A single settings file (settings.ini) defines the database names and access credentials, allowing for a federated database system in the sense that a single web-page can be served by multiple databases which may reside on multiple hosts. This may be of special interest as such an approach would allow us to build to a heterogeneous system of database engines or gain remote access to other database servers. In InsectaCentral's implementation, for example, we deploy Drupal and Chado as core databases and then a SeqFeature::Store database for each of the 200 hosted species. In future versions of InsectaCentral, a laboratory will be able to deploy a local copy of InsectaCentral, a local copy of a Drupal database and connect to the public Chado and SeqFeature::Store databases. They can then deploy their private data as local Chado and SeqFeature::Store databases so that a mix of private data and data from the up-to-date InsectaCentral is seamlessly served to the end-user. Further, the Services module (http://drupal.org/project/services) provides the means for integrating multiple interfaces like XMLRPC, JSON, REST, SOAP, etc avoiding, thus, the need to set up a separate BioMart instance (Smedley et al., 2009). This allows a Drupal site to provide web services to other software via multiple interfaces while using the same callback code. Even though Chado was built to be generic and therefore easy to exchange data between groups, different genome sequencing teams have implemented it in a slightly different way so that cross-communication is not straightforward and adaptors have to be written. The Drupal CMS can become a solution to this compatibility issue between Chado databases.

4.3 Integrating with other software

This generic framework could tap into the concept of bioinformatic work flows, such as those offered by Taverna and Galaxy. This is an interesting possibility to consider and may inspire the EEFG community to use these tools. Meta-servers and software to run bioinformatic applications are constantly being developed. A number of command-line software packages now have their own web-servers and a dedicated journal now exists (the annual Nucleic Acids Research Web Server issue). The most robust and widely used meta-application amongst these is the Galaxy framework. Even though originally developed for genomic data, it has now expanded to other types of data through an active developer community. Galaxy does not offer the main benefits of a CMS (i.e. ease of customization and a rich API). Further, administration of a multi-lab server can be a daunting task for the often over-worked bioinformatician. The biosoftware_bench approach provides full control of the visualization and processing routines. As Drupal is taken up by the GMOD consortium, bioinformaticians who provide new tools would benefit from preparing a biosoftware_bench.inc file (i.e. their software can be easily deployed and laboratories readily can manage and administer it without requiring access to a dedicated bioinformatician).

An increasing number of applications exist for displaying genome data to web-users (e.g. the FlyBase database (Drysdale and Crosby, 2005), the UCSC Genome Browser (Kent et al., 2002), the Ensembl system and the ubiquitous GBrowse (Stein et al., 2002)). As more laboratory groups generate -omic data, there will be a pressing need to develop more such software. One example is the GMODWeb , which builds a website for a Chado database using the Turnkey application (http://turnkey.sourceforge.net). Like GMODWeb, GMOD-DBSF utilizes an external application to drive content deployment but, instead of Turnkey, it uses Drupal, another open-source software. Drupal has the avantages of a broader end-user base and hundreds of developers, and is built to be robust and secure for users and the host server. Because of the large number of functions provided by the core and contributed modules the Drupal solution will become a powerful tool for bioinformaticians.

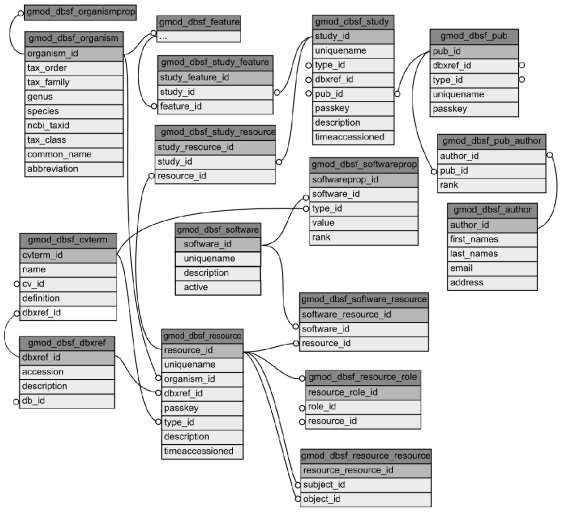
GMOD-DBSF is the only Drupal application built to a generic GMOD API. Another implementation, Tripal, also a GMOD tool (http://gmod.org/wiki/Tripal), is available and in active development. It provides a direct interface with Chado, allowing users to edit the contents of a Chado database. The two modules are not mutually-exclusive as GMOD-DBSF is aimed as a base

module to facilitate development of other modules. With Tripal and the software presented here, the "Drupal solution" provides a feature-set unavailable in any of the other software. Indeed, we could envision multiple Drupal sites linking and sharing their data in a seamless manner.

4.4 Conclusion

The software presented here was built specifically for the research communities that are only now emerging into the -omics era. For example, NGS transcriptome data are widely used to address central biological questions in non-model species but many laboratories do not yet have the means to make the best use of these data. Due to funding constraints, these communities also have a paucity of bioinformaticians. Developed tools must, therefore, be general enough so that they can be used between laboratories but also straightforward to customize so that wet-lab biologists with a bit of training in programming can deploy and maintain the software. Supporting this new cadre of "bioinformaticists" is vital in order for the communities of emerging model species to reap the rewards that NGS technologies have to offer. We have shown that our software can assist with development of bioinformatic web-services. Because Drupal modules are licensed under the GNU Public license and our software was built to be generic and expandable, it would be of interest to the bioinformatic community to expand it. To assist users and developers, we have provided screencast tutorials via another Drupal project, SciVee (http://scivee.tv), which can be accessed via http://gmod.org/gmod-dbsf. We anticipate that the uptake of the Drupal CMS by the bioinformatic community will result in a powerful new set of tools.

Fig. 1. Part of the database schema built by GMOD-DBSF. Chado conventions ensure that this schema can interact with an installed Chado database. Some tables and links omitted for clarity.



5 ACKNOWLEDGEMENTS

We would like to thank the University of Exeter for computational support, drupal.org, InsectaCentral.org users and the Lepidoptera RNAi team for guidance and end-user testing. We would also like to thank Dr Lars Jermiin and three anonymous reviewers for enhancing the quality of the manuscript. No conflicting interests exist. Author contributions: AP conceived, designed and programmed the software, co-ordinated and drafted the manuscript. DH tested the software, advised on design and drafted the manuscript.

Funding: This work was supported by the Max Planck Gesellschaft (AP; DGH), the European Union Research Network GAMEXP (AP) and a Office of the Chief Executive fellowship by the Australian Commonwealth Scientific and Research Organization (CSIRO) to AP.

REFERENCES

Altschul,S.F. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389-3402.

Arnaiz,O. et al. (2006) ParameciumDB: a community resource that integrates the Paramecium tetraurelia genome sequence with genetic data. *Nucleic Acids Research*, **35**, D439 -D444.

Beldade,P. et al. (2008) Butterfly genomics eclosing. *Heredity*, **100**, 150-157.

Collins,F.S. et al. (2003) The Human Genome Project: lessons from large-scale biology. *Science*, **300**, 286-290.

Day,A. et al. (2007) Celsius: a community resource for Affymetrix microarray data. *Genome Biology*, **8**, R112.

Drysdale,R.A. and Crosby,M.A. (2005) FlyBase: genes and gene models. *Nucleic Acids Research*, **33**, D390-395.

Giardine,B. (2005) Galaxy: A platform for interactive large-scale genome analysis. *Genome Research*, **15**, 1451-1455.

Heinicke,S. et al. (2007) The Princeton Protein Orthology Database (P-POD): a comparative genomics analysis tool for biologists. *PLoS One*, **2**.

Hubbard,T. et al. (2002) The Ensembl genome database project. *Nucleic Acids Research*, **30**, 38.

Kent,W. et al. (2002) The human genome browser at UCSC. *Genome Research*, **12**, 996 - 1006.

Lyne,R. et al. (2007) FlyMine: an integrated database for *Drosophila* and *Anopheles* genomics. *Genome Biology*, **8**, R129.

Mungall,C.J. et al. (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, **23**, i337-346.

Ning,Z. et al. (2001) SSAHA: A Fast Search Method for Large DNA Databases. *Genome Research*, **11**, 1725-1729.

O'Connor,B. et al. (2008) GMODWeb: a web framework for the generic model organism database. *Genome Biology*, **9**, R102.

Oinn,T. et al. (2004) Taverna: a tool for the composition and enactment of bioinformatics work flows. *Bioinformatics* **20**, 3045-3054.

Papanicolaou,A. et al. (2009) Next generation transcriptomes for next generation genomes using est2assembly. *BMC Bioinformatics*, **10**, 447.

Schmid,R. and Blaxter,M.L. (2008) annot8r: GO, EC and KEGG annotation of EST datasets. *BMC Bioinformatics*, **9**, 180.

Smedley,D. et al. (2009) BioMart–biological queries made easy. *BMC Genomics*, **10**, 22.

Smith,V. et al. (2009) Scratchpads: a data-publishing framework to build, share and manage information on the diversity of life. *BMC Bioinformatics*, **10**, S6.

Stajich,J.E. et al. (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Research*, **12**, 1611-1618.

Stamatakis,A. et al. (2008) A rapid bootstrap algorithm for the RAxML Web servers. *Systematic Biology*, **57**, 758-771.

Stein,L.D. et al. (2002) The Generic Genome Browser: A Building Block for a Model Organism System Database. *Genome Research*, **12**, 1599-1610.

Terenius O. et al. (in press) RNA interference in Lepidoptera: an overview of successful and unsuccessful studies and implications for experimental design. *Journal of Insect Physiology*

Thain, D. et al. (2005) Distributed Computing in Practice: The Condor Experience. *Concurrency and Computation*, **17**, 323-356.

Yang,Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**, 1586-1591.

Zdobnov,E.M. and Apweiler,R. (2001) InterProScan-an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847-848.

Chapter 5 - InsectaCentral: facilitating comparative genomics with one million insect proteins

This Chapter used the entirety of the thesis to build a unique database system for all Insects. Both the software and the database content are reported. The software is based on the FlyBase Chado database layout and uses the Drupal CMS to manage online content. It is build to be a robust, secure, easy to deploy and species-neutral solution so other laboratories can develop their own Central. The database contains all public insect transcriptome data (from Sanger and Next Generation Sequencing) and a number of secured pre-publication datasets contributed by collaborators.

Citation

Papanicolaou A. and Heckel, D.G.. In preparation for DATABASE (Oxford University Press)

Abstract

We present novel and robust bioinformatic solutions to the transcriptome analysis and dissemination bottleneck. We used the Drupal Content Management System, the GMOD-DBSF module and a new genes4all module to built a repository capable of storing and disseminating transcriptomic data. We deployed an instance called InsectaCentral that houses assemblies and deep annotation of all public insect transcriptomes. The genes4all software is easy to install and freely available for users to build their own -Centrals and is compatible with other bioinformatic modules build for the Drupal Content Management System and offers services such as InterPro, SSAHA2 or JBrowse facilities. All public data are available at http://insectacentral.org. Further, a secure facility for pre-publication data is freely available to researchers wishing to analyse their data through InsectaCentral. It currently holds 214 species with data for 1,489,335 annotated proteins and 5' or 3' untranslated regions. Data from genome-less species have long-term utilities including comparative genomics, functional & biochemical experiments, phylogenetics and molecular ecology.

Introduction

A major benefit of online database resources is that they can bring together researchers to form communities. The centralized availability of data benefits both the researchers and the database resource. The community improves its cohesion and feedback improves the resources. A popular example is annotation consortia supported by genomic databases. The commonality of such endeavors is increasing due to cost-effective genome sequencing. Even with Next Generation Sequencing (NGS), however, genome projects for a large number of species are logistically difficult and current efforts from large sequencing centers produces only low coverage, partly assembled genomes (Bonasio et al. 2010). Until recently, due to its expense, complete transcriptome sequencing was only undertaken after whole genome sequencing was well underway. Genome sequencing consortia are interested in gene models. Gene models, however, can be used not only for annotating whole genome sequencing projects but are also utilized in exploratory research to identify candidates (Pauchet et al. 2009). The attributes of the next generation of genome consortia (not focused on a single species, dispersed worldwide, multi-disciplinary, research grant funded) create a scenario where the consortium has an additional benefit from transcriptome-orientated research: they can be used to form annotation groups long before whole genome assemblies became available. In previous work (Papanicolaou et al. 2008), we showed how bioinformatic support for Expressed Sequence Tag (EST) and transcriptome sequencing of a wide research community can benefit both functional and evolutionary biologists. Indeed with NGS, community support is of

increased importance: laboratories with small budgets can now generate transcriptomes and gene models of their favorite species but are often unable to process, mine and disseminate their own data and findings. NGS has indeed removed the sequencing bottleneck which hampered sequence-based projects. With the current low cost of complete transcriptome sequencing (Wang, Gerstein, and Snyder 2008; Papanicolaou et al. 2009; Ferguson et al. 2010), it is possible to complete transcriptome sequencing before or during whole genome sequencing. Further, transcriptomes of an increasing number of species are being released (Papanicolaou et al. 2009; Ferguson et al. 2010; Kang et al. 2004; O'Neil et al. 2010; Beldade et al. 2006) and often each group needs to generate their own database for each species. In is not unexpected, therefore, that the next bottleneck is the analysis and dissemination of this vast amount of data. This has also provided a whole new need for bioinformaticians in research fields where they were not previously needed. The bioinformatic community is assisting in creating resources that can harvest these data and provide immediate benefits to researchers. For most laboratories, one of the immediate needs is the deployment of an analysis infrastructure, driven by one or more underlying databases. This can lead to a proliferation of disparate resources and the lack of a standardized approach impacts on quality. Our own work falls in this category (Papanicolaou et al. 2008) but some resources have been proven to be of such wide-utility (e.g. ButterflyBase has been cited 23 times in the period of January 2008 to October 2010; source ISI Thompson Reuters accessed 3rd of October 2010) that supporting these community resources is at least as important as producing the raw data in the first place. Development of bioinformatic tools, however, should not be undertaken in solitude for reasons of efficiency and standardization. Currently no consortium has shown interest in building a species-neutral platform despite an increasing number of genomes being published. The data dissemination bottleneck may lead to a proliferation of costly, custom, non-standardized in-house solutions.

Here we present a robust solution to this problem. We utilize open-source, standard tools part of the General Model Organism Database (GMOD) consortium that builds on previous work (Papanicolaou and Heckel 2010) on the GMOD and the Drupal Content Management System (CMS) toolkits, two open development platforms for genomics and online content respectively. The resulting resource, genes4all, utilizes the Chado database to provide reference transcriptomes, their annotation and the ability to curate it. We then used est2assembly (Papanicolaou et al. 2009) to generate reference transcriptomes for all insect species found in GenBank and the Short Read Archive of NCBI (http://www.ncbi.nlm.nih.gov/sra). Further, we provided a secure facility for storing pre-publication NGS datasets of collaborators leading to InsectaCentral currently holding 214 species with data for 1,489,335 annotated proteins and UnTranslated Regions (UTR).

Materials and methods

Genes4all & InsectaCentral specifications

Genes4all's main aims descend from ButterflyBase and its supporting software, PartiGene (Parkinson et al. 2004). The specifications and implementation have been, however, reconsidered. Via the est2assembly software, we still generate a reference transcriptome using assemblies of EST data and store both the reference and the constituent reads in a relational database. Additional to the traditional capillary (Sanger) sequencing est2assembly also supports the 454 technology. The Illumina and SOLiD technologies are not supported yet due to the large amounts of raw data. Unlike PartiGene, we focused on utilizing much of the GMOD framework: the BioPerl library (Stajich et al. 2002); the Chado (Mungall and Emmert 2007) and Bio::SeqFeature::Store (SeqFeature) database schemas; the GBrowse sequence viewer (Stein et al. 2002) and JBrowse (Skinner et al. 2009) . In order to develop and deploy in a modular fashion, we have relied on the Drupal CMS which also handles tasks such as visualization, user authentication and a module's Application Programmatic Interface (API) exposure. Further, we used Drupal's *tabs*, *gmod_dbsf* and *biosoftware_bench*. These modules to handle tabular content, Chado/GMOD integration and deploy bioinformatic software applications respectively (Papanicolaou and Heckel 2010).

InsectaCentral data processing

The data are either derived from public data (NCBI's dbEST and the Short Read Archive) or are pre-publication NGS collections from collaborators. They are processed with est2assembly to produce an assembly of contigs, predicted proteins and annotations. Briefly, sequences were downloaded from EBI and preprocessed to remove any vector, contaminant and adaptor sequences. Sequences meeting the quality and length criteria were clustered using MIRA2.9.37 (Chevreux, Wetter, and Suhai 1999) and Newbler2 (454 Life Sciences) with varying parameters before choosing the optimal one based on number of reference proteins identified and coverage of reference proteins. In each case, MIRA out-performed Newbler and thus one of the MIRA assemblies was used. In the case of Sanger only sequences, we utilized the trim_assembly -debris function in order to include singleton ESTs. In 454-only datasets, we significantly reduced the computational power needed by using trim_assembly without the -debris option since no singletons should exist in data with hundreds of thousands of ESTs. In both cases, we predicted proteins using prot4EST 2 (Wasmuth and M. L. Blaxter 2004) before performing deep-annotation. The prot4EST program has four tiers of prediction from most accurate to least: similarity to known protein, ESTScan, Decoder and longest ORF from a six-frame translation. Proteins derived from the latter

are often erroneous since even a UTR region can hold a short sequence between what appears to be a start and a stop codon. A high-performance computing cluster (HECToR) and our local Condor-driven PC farms at Exeter and Canberra allowed for deep annotations including InterProScan domains (Zdobnov and Apweiler 2001), Gene Ontology (GO) (Ashburner et al. 2000), Enzyme Classification (EC) (Bairoch 2000) and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto 2000) using annot8r (Schmid and Blaxter 2008) and BLASTX with a bit-score cutoff of 60 bits. The KOG data were derived from NCBI and Insect-specific terms were captured from the *Drosophila melanogaster* data using BLASTX with a bit-score cutoff of 60 bits (*ca* 1e-12 evalue cutoff).

Data warehousing

Using est2assembly GFF3 files of the assembly, protein predictions and the annotations are produced using unique identifiers according to the est2assembly schema. For example, IC7144AaEcon124 is composed of i) a two letter database ID (IC for InsectaCentral in this case); ii) the NCBI taxonomy ID of the species (7144); iii) an assembly version (Aa being first, Ab second etc); iv) the data type ID; v) a serial number (124). Each object identifier is unique and permanent. Unlike its precursor, PartiGene, new assemblies can exist alongside old ones and there is no restriction of one contig object - one protein object as the serial numbers are not shared between data types. The only exception is between (automated) predicted Open Reading Frames (ORF) and the translations into proteins which have a natural one-to-one relationship. Finally, all the data are stored in GMOD-compatible database schemas: Chado, and optionally the SeqFeature::Store.

As most GMOD members are using genomes as a reference, however, a new approach of implementing Chado had to be developed. Details available in (Papanicolaou et al. 2009). Briefly, est2assembly utilizes three reference objects for sequence data (called 'features' in Chado-speak): contigs as cluster of ESTs (with the data type key being Econ); the predicted Open Reading Frame (Aorf) and the predicted protein (i.e. polypeptide; Apep). Each reference object allows for anchoring constituent sequences (e.g. reads forming a contig) and a number of annotations. The use of placeholders allows for cross-referencing between each reference object and ontologies are used to anchor annotations on the references. Considering that we are populating with whole assemblies of hundreds of species, the resulting database scales unexpectedly well. Compex queries, however, had to be simplified via the use of materialized views. Such views are pre-computed complex queries, the results of which are stored in a new table. A minor disadvantage is that with every new release, all materialized views must be computed, an exercise which delays the release for several

hours.

Data visualization

Once a Chado database is built with specific content, a -Central can be deployed to facilitate visualization. Unlike the majority of Chado users, genes4all is a dedicated transcriptome database software using Chado without a reference genome. We had, therefore, to innovate a visualization interface. We used the Drupal CMS to build a meta-module, i.e. a module with a number of interconnected yet optional sub-modules. Specifically, genes4all has an 'explore', a 'curate', a 'download' and a 'experiment' module. Each of these modules uses the gmod_dbsf Drupal library (Papanicolaou and Heckel 2010) which has standard functions for manipulating a Chado database. These modules are aimed to be read-only and not manipulate Chado. The genes4all_curate module, on the other hand, provides functions to write to a Drupal database and synchronize it with Chado. The Drupal CMS provides user authentication in order to ensure that only priviliged users (e.g. curators) can write to the Drupal database and only authorized users (e.g. head curators) can perform a write operation to Chado. We utilized Jquery and the Dojo Toolkit, two open-source JavaScript libraries, to deploy interactive content.

Further, we provided interfaces for GBrowse and JBrowse, two standard browsers that allow users to explore reference objects in a familiar fashion. Chado is, however, a highly normalized database and we found that the large number of references and annotations is unacceptable for high-throughput visualization tools such as GBrowse. For that reason, genes4all allows the use of SeqFeature::Store, a normalized schema. To assist with queries, we recommend that every species has its own database. We find that this solution is cumbersome with databases housing a large number of number of species and provide, therefore, as an option an alternative approach via the JBrowse interface. This method retrieves data from the Chado materialized views and produces temporary JavaScript Object Notation (JSON) files which are subsequently used by JBrowse. As new software are released frequently (e.g. a JavaScript version of the Artemis annotation tool), software developers may contact us if they wish for genes4all to provide support for their program. The current implementation of GBrowse is 1.70 as that was the stable version during development and we are using the 1.1 release of Jbrowse. For InsectaCentral, due to the large amounts of data, makes use only of Jbrowse, foregoing thus the need for a SeqFeature::Store schema.

Results

Content

InsectaCentral currently contains data from the NCBI (dbEST and Short Read Archive) and pre-publication 454 pyrosequencing data provided directly by scientists. A total of 1,489,335 annotated proteins and UTR regions (i.e. contigs) are available from 214 species. Some species are more data-rich than others as shown by Figure 1. As more scientists make their 454 pyrosequencing data available to InsectaCentral, or when we initiate the processing of Illumina RNA-Seq data, the distribution will shift to the right but InsectaCentral. For each contig we have enforced a protein prediction in order to capture proteins not identified in model species. A significant number of those, however, is likely to be UTR: Sanger-capillary data are often not sufficiently rich to link CDS with potentially long UTRs. Traditionally this has been resolved with full-length cDNA sequencing but the new Illumina RNA-Seq technology is a better alternative. Due to the large amounts of raw data, InsectaCentral does not process Illumina RNA-Seq data currently but it is an area we are actively working on.

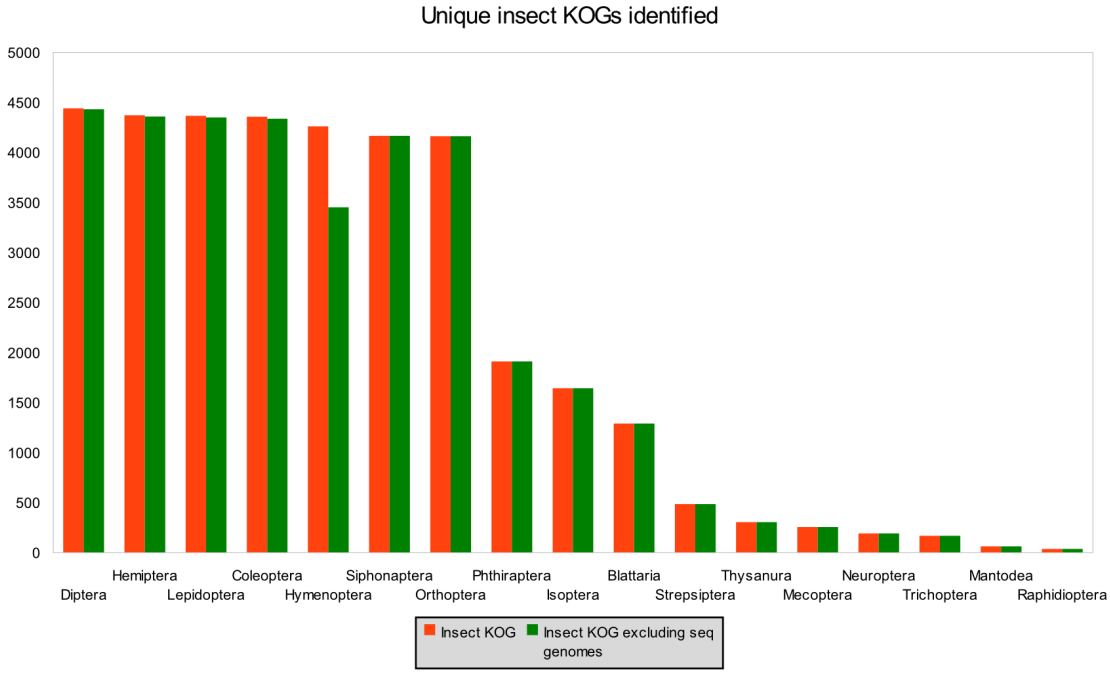


***Figure 2**: Unique insect KOG (euKaryotic clusters of Orthologous Genes) terms identified including species with a published genome project (red; left) or excluding them (green; right). Because D. melanogaster had been used to classify KOGs, the Diptera clade can be assumed as the saturation point. The green bars for Hemiptera, Lepidoptera, Coleoptera, Siphonaptera and Orthoptera have approached this saturation point meaning that transcriptomes as processed by est2assembly and presented by InsectaCentral are sufficient to identify all the KOGs within each of these clades. Some clades, such as the Lepidoptera, are species-rich allowing for within clade comparisons and enrichment of the KOG ontology. Other clades such as Siphonaptera and Orthoptera are represented by a single species (the oriental rat-flea Xenopsylla cheopis and the cricket Gryllus campestris – two private data sets produced by one and two 454 transcriptome runs respectively) but they too will be essential for expanding the insect KOGs.*

Scientists have been making use of the NGS technologies to capture transcriptome sequences from species without a genome and InsectaCentral is the only available platform where these data can be easily captured and presented. InsectaCentral already comprises the most data-rich resource for insect gene-finding because it is not specific to a specific order of insects. We identified the proportion of insect KOGs in the database (Figure 2) and 7 orders are already represented with all the KOGs identified from *D. melanogaster*. Four of these orders have a published genome sequence for at least one reference species but as the figure shows, if one takes into account only

transcriptome data from non-sequenced species, coverage is maintained. Further, two of these orders are not species rich: Siphonaptera and Orthoptera are represented by one species each via a 454 pyrosequencing dataset. Our ability to include pre-publication data allows us to have a more accurate overview of available -omic data. In addition, once these data are published, scientists can authorize us to make the data public: thanks to the genes4all software, this operation is instaneous and seamless.The Drupal genes4all module for visualization

Exploration

The genes4all_explore module allows users to search for data from one or more species using a phylogenetically aware 'checkbox-tree' which allows whole families or orders to be selected (Figure 3). Access control is enforced by Drupal's user-management and can be set so that public data will be open to any user and no registration is needed (i.e. 'anonymous' users are given access). Administrators can set certain species or certain features to be accessible by certain users or by certain roles. Thus users with private data can log in with their account details in order to be able to select their secured species or dataset of interest. Regardless, once a species is selected, then the following queries are limited by organism. Users can query for contigs based on any of the pre-computed annotations (BLAST description line, GO, KEGG, EC or InterProScan findings). For datasets where this information is available (i.e. dbEST or submission from an individual contributor), it is possible to limit a search by characteristics of the cDNA library. They can also mine for contigs containing Single Nucleotide Polymorphisms (SNPs) markers. In the future, we hope to be able to integrate genome data and we envision linking contigs with these markers to genomic regions. When the user has decided on a set of query characteristics, they can mine for objects which meet 'all' or 'any' of these criteria. When a query is made, users are presented with a table of objects which meet the query specifications. Users can then use the bookmark button to store their query as a bookmark for when new data becomes available; select one or more of the objects and download them in either FASTA or GFF3 format; or click on the object to go to a summary page.

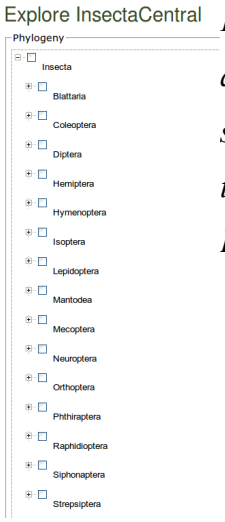


***Figure 3**: More than 200 species are made available for searching. Users can thus select to search multiple species of interest and not limit themselves to published genomes or unclustered EST data.*

Summary pages

Each data type (a.k.a. feature) has its own type of summary page but as they are implemented in a similar fashion, we give an overview of the contig summary pages (i.e. genes). We've implemented the popular format of gene pages, inspired by other large genome databases (Howe et al. 2008). The first tab of the gene page gives an overview of how this contig was constructed, provides links to the cDNA library and the associated ORF and protein pages. Further, links to JBrowse allow users to explore how the feature was generated and what annotations it holds (Figure 4). The second gene page tab provides an array of electronic annotations which have been transferred to this object such as GO terms or words from BLAST description lines. The third tab, provides JavaScript graphs of the coverage (in case of contigs only) and BLAST annotation hits (Figure 5). During manual curation, we find these graphs useful for determining untranslated regions (UTRs) or chimeric reads: UTRs have no BLAST similarity to proteins (but may have to genomes) and chimeric reads show a trough where the join is occurring (Lee et al. 2007). Finally, other tabs hold the sequence data in FASTA and GFF3 format respectively, allowing users to download not only the sequence but also the assembly of a contig.

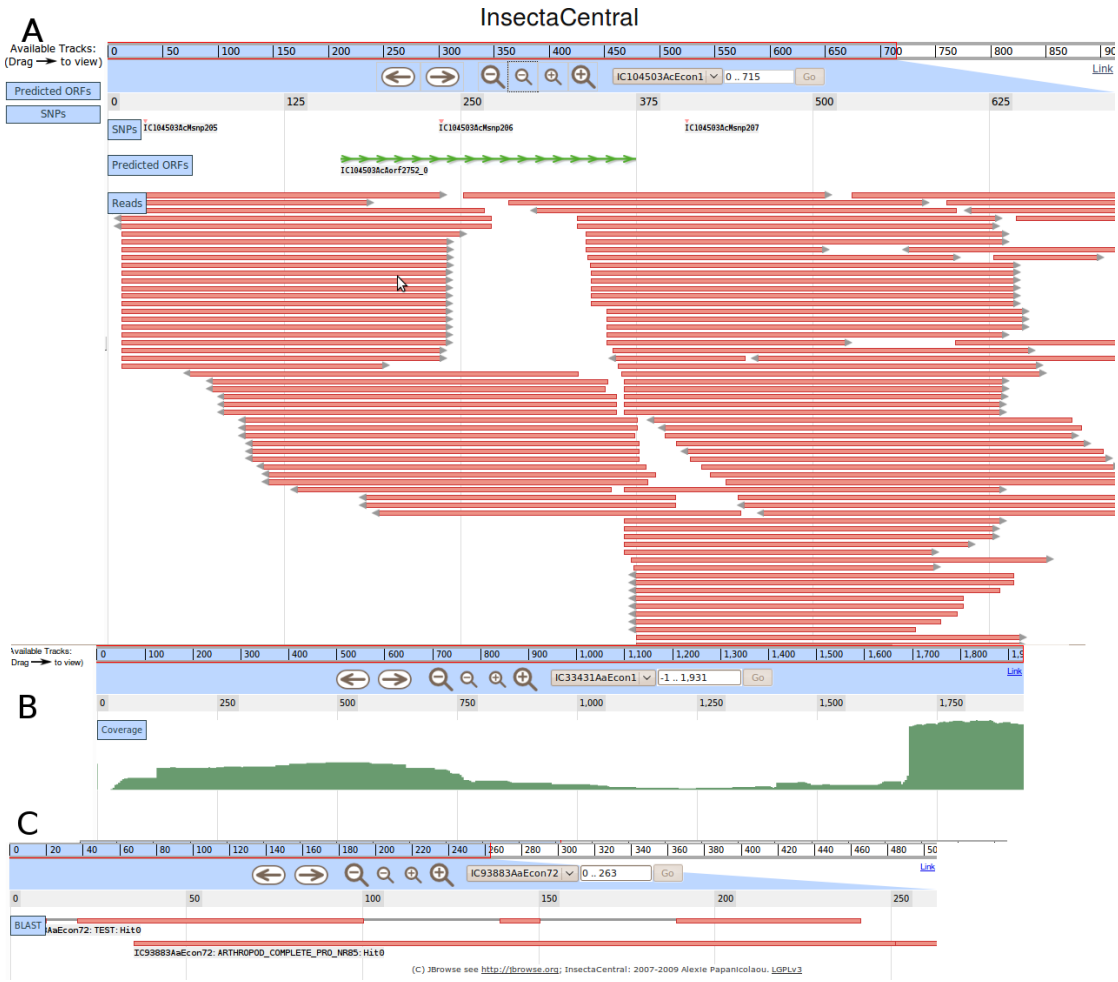


Figure 4: *The basis of InsectaCentral is to make data available so that scientists can make their own conclusions. One approach is to explore features interactively using JBrowse, a JavaScript application similar to the GBrowse software. A) The predicted Open reading frames allow users to see exactly where the ORF lies. B) Tracks such as coverage can be informative to determine the structure of the gene and the relevant expression levels of exons. C) By databasing annotations such as the BLAST similarity results complements the predicted ORF and users can see the support for particular annotations.*

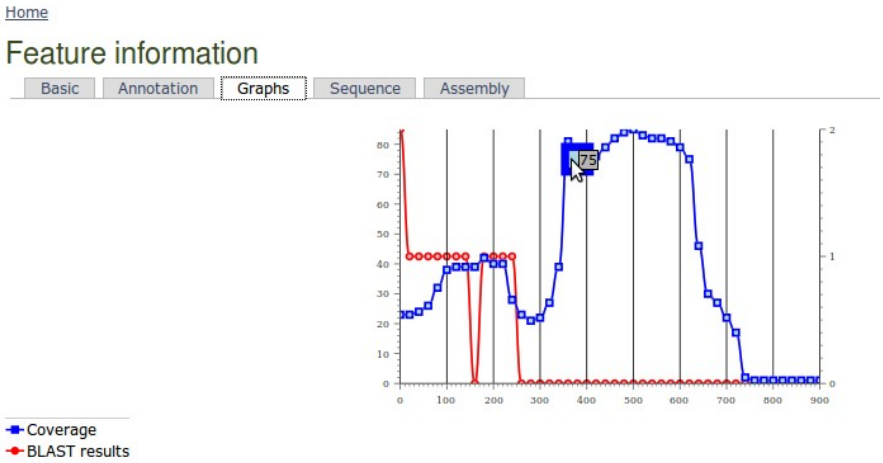
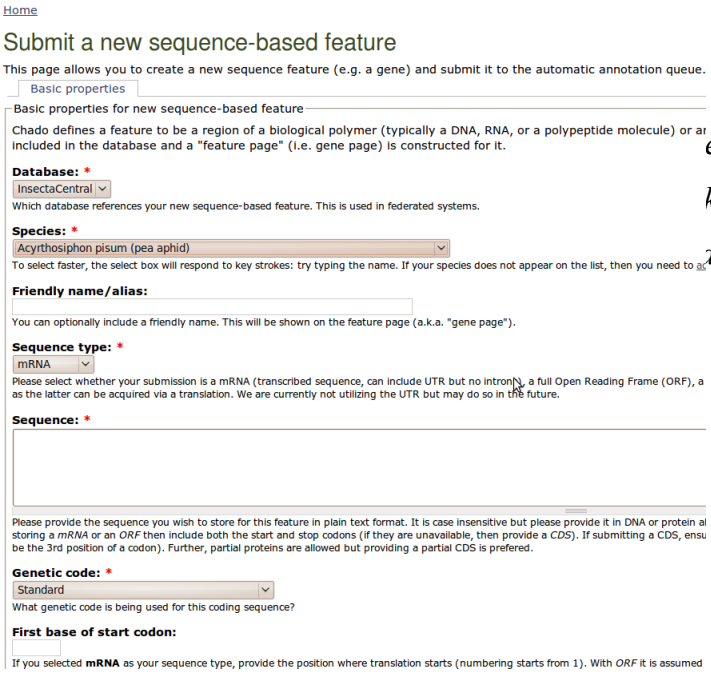


***Figure 5**: Visual access to the coverage, depth and annotations allow biologists to better comprehend the biological importance of any specific feature. For example, this contig has a region in the middle which is highly transcribed but has no similarity to any known protein but the left-most region does have significant similarity to 1 known protein. In depth investigations show this is a contig composed of 3' UTR which failed to assemble to the rest of the CDS. In Lepidoptera, the 3'UTR tend to be rather long and the specific library was composed from samples with high heterozygosity (i.e. multiple outbred individuals).*

Browsing reference sequences

Further links to the GBrowse or JBrowse interface are available from the summary pages. Using these browsers, users can make an in-depth investigation of how the feature was built (via the EST track), any annotations that are available (via the BLAST similarity or SNP tracks). An ORF track allows users to migrate to the ORF reference sequence and then to the predicted peptide. Where available, BLAST annotation tracks have links of the BLAST hit to external databases such as the UniProt Knowledgebase (Suzek et al. 2007). For GBrowse, we have also used the HapMap approach of viewing SNP data by providing a pie-chart of allele frequencies derived from the EST reads forming the contig. This may not be an optimal approach as only the assembly is considered. The concept of re-aligning all reads to the reference before determining frequencies is being considered for future releases.

Curation module


Home

Submit a new sequence-based feature

This page allows you to create a new sequence feature (e.g. a gene) and submit it to the automatic annotation queue.

Basic properties

Basic properties for new sequence-based feature

Chado defines a feature to be a region of a biological polymer (typically a DNA, RNA, or a polypeptide molecule) or ar
included in the database and a "feature page" (i.e. gene page) is constructed for it.

Database: *

InsectaCentral

Which database references your new sequence-based feature. This is used in federated systems.

Species: *

Acyrthosiphon pisum (pea aphid)

To select faster, the select box will respond to key strokes: try typing the name. If your species does not appear on the list, then you need to ac

Friendly name/alias:

You can optionally include a friendly name. This will be shown on the feature page (a.k.a. "gene page").

Sequence type: *

mRNA

Please select whether your submission is a mRNA (transcribed sequence, can include UTR but no introns), a full Open Reading Frame (ORF), a
as the latter can be acquired via a translation. We are currently not utilizing the UTR but may do so in the future.

Sequence: *

Please provide the sequence you wish to store for this feature in plain text format. It is case insensitive but please provide it in DNA or protein al
storing a *mRNA* or an *ORF* then include both the start and stop codons (if they are unavailable, then provide a *CDS*). If submitting a CDS, ensu
be the 3rd position of a codon). Further, partial proteins are allowed but providing a partial CDS is prefered.

Genetic code: *

Standard

What genetic code is being used for this coding sequence?

First base of start codon:

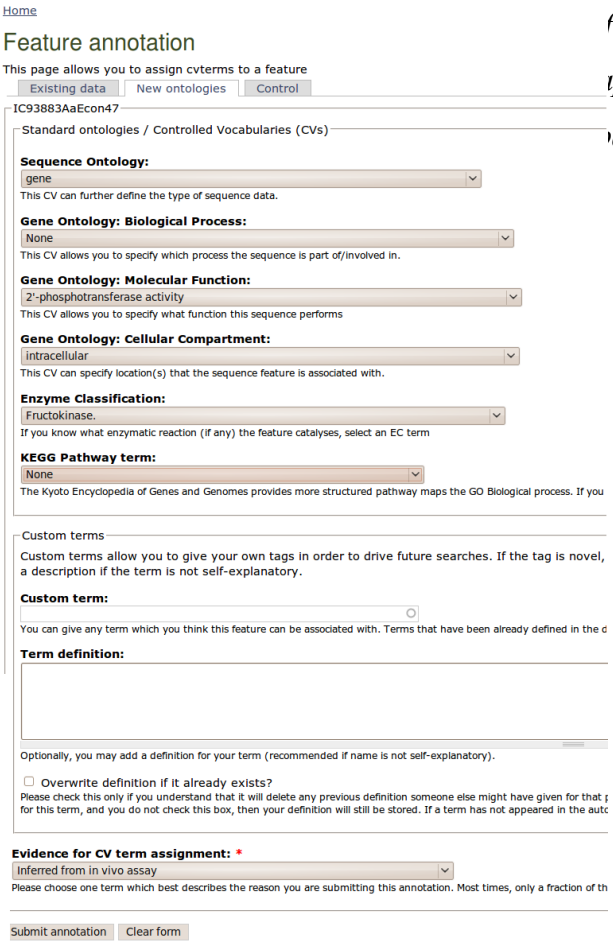
If you selected **mRNA** as your sequence type, provide the position where translation starts (numbering starts from 1). With *ORF* it is assumed

ed by authorized users. These enter

ke any other InsectaCentral feature

ame is linked to the feature).

Linked to each gene page is the ability for authorized users to submit new Open Reading Frames related to a contig (Figure 6). The community, therefore, can play the vital role of editing automated predictions. First the user finds the contig of interest and initiates a curation protocol by clicking 'Curate this!'. Then they can specify which other contigs, if any, they think should be assembled with this contig and if the gene uses an alternative codon table (such as the mitochondrial one). A consensus is generated for them which they can utilize to select start and stop points for an ORF or they can type the sequence themselves. In either case, the proposed ORF is evaluated and quality checked (start/stop codons, length and codon usage) and the user is asked to verify and sign their submission. A new Corf (curated ORF) and Cpep (curated peptide) object is then created using the latest assembly version from the features used (i.e. which assembly the user had access to in order to make their curation). These curated objects then enter an annotation queue and summary pages can be generated. An administrative page allows a selected set of individual head-curators to approve these submissions and a tag that they have been curated along with the name of the curator appears on the Corf/Cpep gene pages. In a similar fashion, curators can link any feature with one or more terms from ontologies (such as the GO or EC) or provide custom terms (Figure 7). This functionality allows InsectaCentral to serve as a community annotation system.

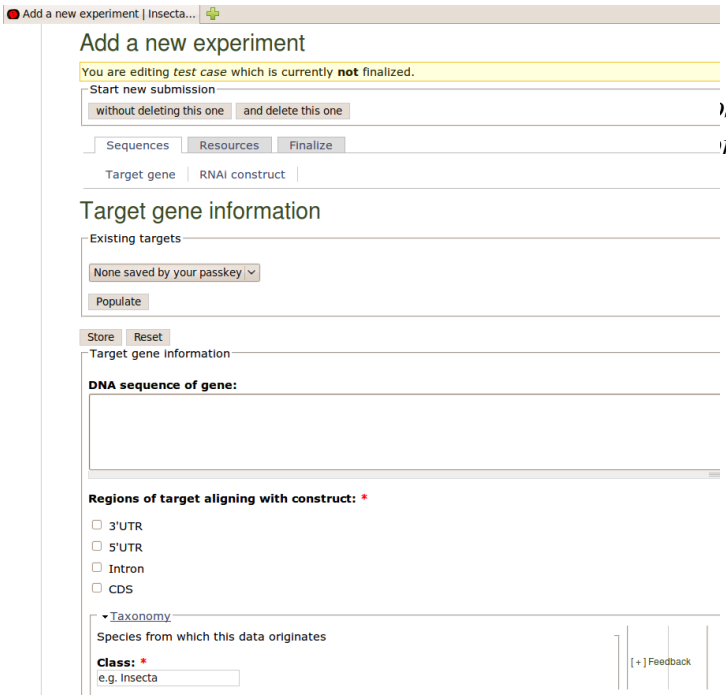


functional annotation to a feature. These approved (by a head-curator) become ages.

Experiment module

InsectaCentral's central aim is a community tool. We wrote an experiment databasing module using the Drupal API which was made available previously (Papanicolaou and Heckel 2010). An extension of the Chado schema allows us to handle studies and resources using Controlled Vocabularies. The Lepidoptera RNAi group has piloted the service with a study aimed at understanding when RNAi silencing can work in butterflies and moths. It is available from http://www.insectacentral.org/RNAi and negative results are of particular interest as they are usually not published. An important feature is that the submitted experimental data is not available to the public but only to the author and authorized individuals (e.g. the Lepidoptera RNAi group). The module offers the ability to store one or more studies and lock them with a private passkey which can then be used by the author or their research group to revisit them if a change or information is needed. For the RNAi work, we provide six sets of resources which need to be databased: the target gene and RNAi construct (stored in Chado's feature table); experimental animals; delivery and assay protocols (stored in the new resource table which closely mimics the feature table; Figure 8). Further, a publication title can be optionally provided but a communicating author is needed for

handling responsibility of the submission. Four of these sets (publication, target gene, construct and experimental animals) allow for database cross-referencing such as Pubmed, GenBank or a stock center. Due to the standardized API and use of Controlled Vocabularies, it is straightforward to offer similar services for other experiments, depending on community feedback.



ssible via a web interface. A security
ns to be made by the submitters or

Similarity searching

A common method for identifying genes of interests in any sequence database is to use a BLAST or other similarity search server. Genes4all integrates well with another Drupal module, Drupal Bioinformatic Software Bench (biosoftware_bench). Currently, available tools include the BLASTALL (Altschul et al. 1990), SSAHA (Ning, Cox, and Mullikin 2001), InterProScan and annot8r software. Briefly, users can identify genes of interest, download them or link to the genes' summary page. The biosoftware_bench module provides many improvements over traditional BLAST servers: it allows users to perform queries with multiple BLAST algorithms, submit large multi-FASTA queries and it makes use of our in-house distributed computing facility. Further, it can be used for non-local data: it provides the facility to do searches against non-transcriptome datasets that the administrator judges to be of use to the user community, such as genomes and popular non-redundant protein sets, or users can upload their own subject databases for their personal use. The resulting search data are presented in a table with a graphical overview and using BioPerl it offers a variety of BLAST output formats.

Discussion

InsectaCentral content and utility

We used genes4all and insect transcript data to build the first taxon-wide, gene-focused GMOD database. InsectaCentral holds the transcriptomes of 194 insects, including pre-publication data from collaborators. The datasets of many of those are poor: only 94 of these species have more than 1,000 gene objects. One of the main reasons for paucity of such data is not because they have not been produced but because they have not been deposited in one of the public databases. In an attempt to entice deposition but also because we are interested in providing a community service, we had to consider supporting private datasets. A number of laboratories produce large sets of transcriptome data (especially since NGS became widely available) but do not opt for a public release even though only a fraction of the data is of interest to them: the overhead, i.e. the amount of work needed, is higher than any perceived benefits. InsectaCentral offers, therefore, the capability for registered users to agree to upload their pre-publication data on a secure section of InsectaCentral and we process, annotate and make it available to them. At a later date, they can opt to make all the data public. In the meantime, we can provide multiple users with group acccess so that they can enhance their laboratory's data-mining capability. Currently (October 2010), six such laboratories have opted-in and, as “beta-testers”, they have contributed on the design of the offered facilities.

Regardless of origin, each dataset undergoes deep-annotation: users can use annotations and predictions or use cDNA library characteristics to mine for genes of interest. The collection of all these data under one roof, allows or a one-stop solution for biologists working with molecular data of insects and feel that FlyBase (Wilson, Goodman, and Strelets 2007) and VectorBase (Lawson et al. 2007) is not meeting their needs. Importantly, the web-interface has been built with the help of small group of wet-lab biologists in an attempt to understand how they mine for information and how should it be best presented to them. Future versions should extend this beta-testing group and improve the interface.

Due to the species richness of the resource, comparative genomics questions can be placed in a phylogenetic context which is limited only by the availability of public data. Such investigations can determine novel genes families, provide putative functions, and survey for regions which might be selected in one or more clades. As the resource is currently focused on transcriptomes, homologous multi-species UTR sites or intron read-throughs can be to look for signatures of selection at non-coding DNA. We envision that the curation of gene models can be utilized by the community to include such features. Moreover, the generation of reference gene models for a

particular gene family across multiple species can provide novel phylogenetic markers to investigate the unresolved relationships between and within Insecta orders.

A moving target

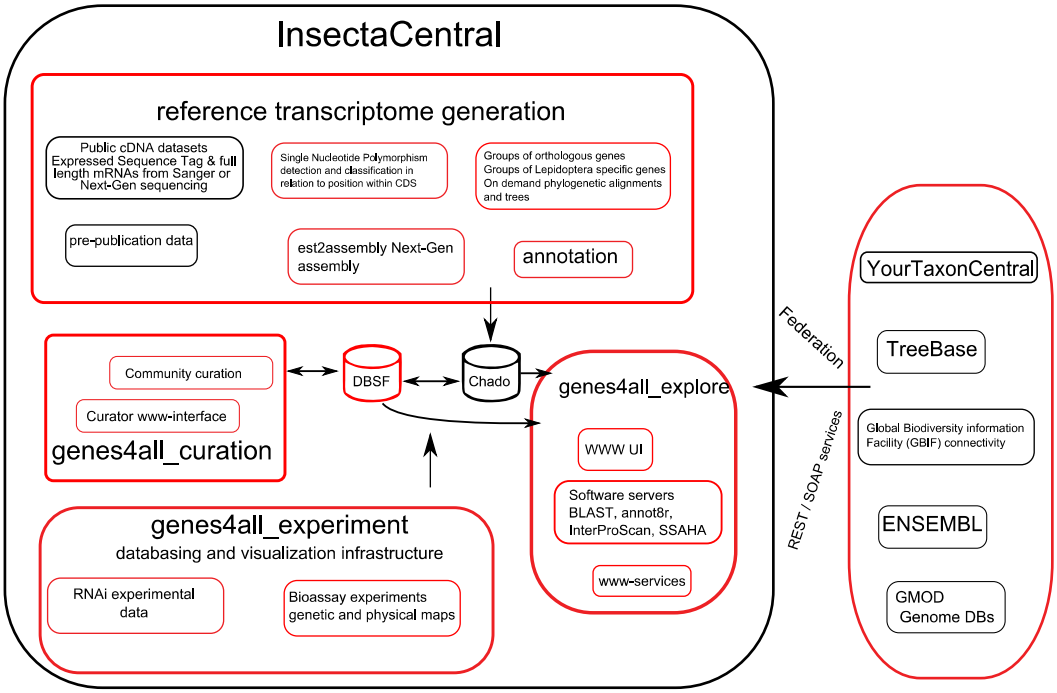


***Figure 9**: The big picture: InsectaCentral is the data analysis and dissemination platform initially developed in concert with the Lepidopteran community. Red boxes represent work undertaken by us and black boxes represent existing work or via collaborations. The right-hand side (federation) is currently work in progress.*

The diversity of data-types is rapidly increasing. Federation with other databases and a comparative module for InsectaCentral is lacking but planned (Figure 9). Comparative studies, even within one species, can be undertaken by considering data produced from different tissues and/or different laboratories. Further information can be gained via microarray approaches if they comply to MIAME, the Minimum Information Criteria for Microarray Experiments. Currently, such resources are not widely available but with the invention of Illumina and RNA-seq (Wang, Gerstein, and Snyder 2008) this situation may change. Experiments utilizing this affordable technology can provide valuable information via transcriptome profiling. Currently, these experiments are only available via the standard literature. InsectaCentral was designed prior to RNA-Seq data. Support for Illumina RNA-seq is planned but would require a rethinking of the warehousing strategy. This would allow us to provide dense digital expression profiles for each contig. In addition, further

future work is concentrated in supporting reference genomes so that a reference genome can be used to anchor non-model species transcriptomes. Finally, ecological data in relation to population variants are expected to become widely available thanks to inexpensive Illumina or array-based technologies. Despite the current deficits, InsectaCentral provides the a develop-friendly open-platform. We used est2assembly for transcriptome assembly and annotation but any assembly platform could have been used: our ultimate aim is to provide a system which is designed for emerging models, is open and is GMOD-compatible so that laboratories can integrate with their current data as seamlessly as possible.

A community resource

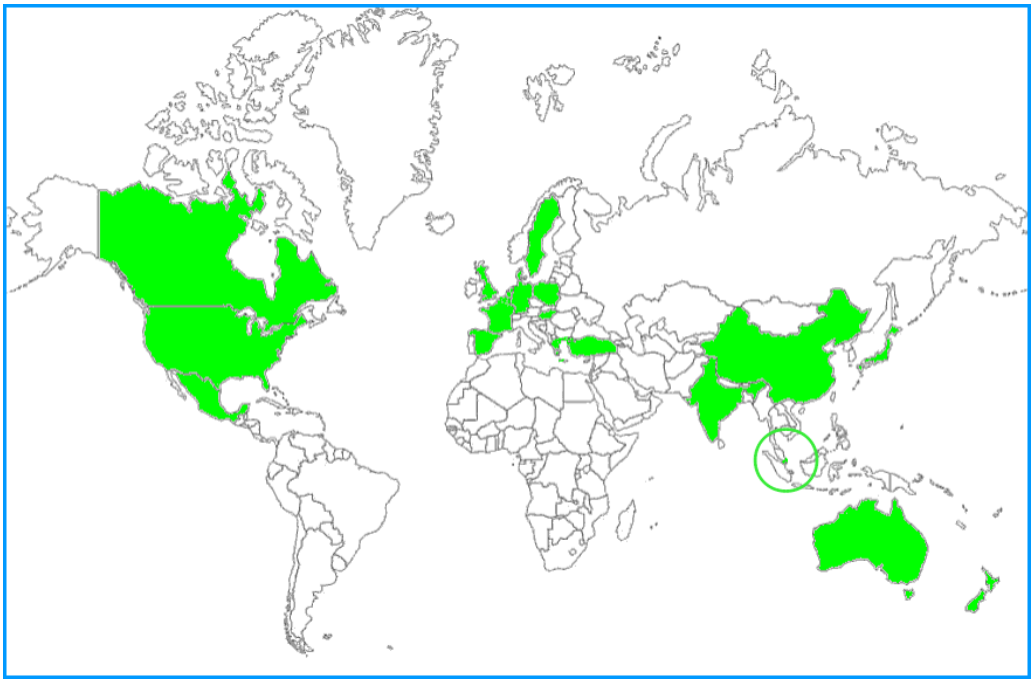


***Figure 10**: Countries participating in the first phase of the RNAi experiment database (http://insectacentral.org/RNAi). This world-wide effort was the first attempt to identify the factors influencing the feasibility of RNAi studies on Lepidoptera.*

This plethora of gene models, especially from species without a sequenced genome, has one significant effect that was discovered during our original ButterflyBase work: online resources do bring research communities together. InsectaCentral has fostered working groups like the RNAi group (Terenius et al 2010; Figure 10). By providing a user-friendly software, a pre-publication facility and deep-annotation, we can entice the sharing of data. The aim of InsectaCentral is to convince researchers to enhance the transcriptome resources for Insects (and improve the distribution in figures 1 and 2): non-genomics wet-lab biologists who feel that 90% of the EST data

they have produced are useful to others will now have no bioinfomatic obstacle for sharing them. Indeed, this sharing can be instrumental in the formation of communities prior to full genome sequencing. The above bioinformatic innovations are part of a general shift of the community towards collaborative bioinformatics utilizing stricter standards, species-neutral solutions and open-access frameworks. Further, via InsectaCentral, we have designed a resource which serves insects species without a sequenced genome, improves the standards of transcriptome sequencing, reporting and provides a platform where nascent insect consortia can form. Indeed, being species-neutral any number of -Centrals could be built and interface with each other. As more transcriptomes are sequenced and analysed we will have a valuable resource for mining taxon or species-specific proteins. We trust that these rapid advances in transcriptome analysis, and the bioinformatic bottleneck they will produce, will benefit notably from genes4all and InsectaCentral.

Data submission and access statement

All publicly available data on InsectaCentral are freely accessible without registration. They are released under the terms of Limited GPL v3 and can therefore be used for academic or commercial reasons without restrictions as long as discoveries and derivate works cite this article. For users to included in data in the public section of InsectaCentral, they should first submit their raw data to the TraceArchive or Short Read Archive (NCBI) and notify us. We can offer assistance of this step for laboratories without bioinformatic expertise. users wishing to take advantage of the InsectaCentral platform for their private datasets, should contact the communicating author in first instance. Our goal for the future is to develop the project guided by the community. Therefore, we welcome requests and contributions.

References

Altschul, S., W. Gish, W. Miller, E. Myers, and D. Lipman. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215: 403 - 410.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, and J. T. Eppig. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* 25, no. 1: 25.

Bairoch, A. 2000. The ENZYME database in 2000. *Nucleic Acids Research* 28, no. 1 (January): 304-5.

Beldade, P., S. Rudd, J. D. Gruber, and A. D. Long. 2006. A wing expressed sequence tag resource for Bicyclus anynana butterflies, an evo-devo model. *BMC Genomics* 7: 130.

Bonasio, Roberto, G Zhang, Chaoyang Ye, Ns Mutti, Xiaodong Fang, and Nan Qin, G. 2010.

Genomic Comparison of the Ants Camponotus floridanus and Harpegnathos saltator. *Science* 329, no. 5995 (August): 1068-1071. doi:10.1126/science.1192428.

Chevreux, B., T. Wetter, and S. Suhai. 1999. Genome sequence assembly using trace signals and additional sequence information. In *German Conference on Bioinformatics*, 45-56. Citeseer.

Ferguson, L., S. F. Lee, N. Chamberlain, N. Nadeau, M. Joron, S. Baxter, P. Wilkinson, A. Papanicolaou, S. Kumar, and T. J. Kee. 2010. Characterization of a hotspot for mimicry: Assembly of a butterfly wing transcriptome to genomic sequence at the HmYb/Sb locus. *Molecular Ecology* 19, no. s1: 240-254.

Howe, A. D, M. Costanzo, P. Fey, T. Gojobori, L. Hannick, W. Hide, D. P Hill, R. Kania, M. Schaeffer, and S. St Pierre, others. 2008. Big data: the future of biocuration. *Nature* 455, no. 7209: 47.

Kanehisa, M., and S. Goto. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* 28, no. 1: 27 - 30.

Kang, L., X. Y Chen, Y. Zhou, B. W Liu, W. Zheng, R. Q Li, J. Wang, and J. Yu. 2004. The analysis of large-scale gene expression correlated to the phase changes of the migratory locust. *Proceedings of the National Academy of Sciences of the United States of America* 101, no. 51: 17611.

Lawson, D., P. Arensburger, P. Atkinson, N. J. Besansky, R. V. Bruggner, R. Butler, K. S. Campbell, G. K. Christophides, S. Christley, and E. Dialynas. 2007. VectorBase: a home for invertebrate vectors of human pathogens. *Nucleic Acids Research* 35, no. Database issue: D503 - 505.

Lee, Byungwook, Taehui Hong, Sang Jin Byun, Taeha Woo, and Yoon Jeong Choi. 2007. ESTpass: a web-based server for processing and annotating expressed sequence tag (EST) sequences. *Nucleic Acids Research* 35, no. Web Server issue (July): W159-62. doi:10.1093/nar/gkm369.

Mungall, C. J., and D. B. Emmert. 2007. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics* 23, no. 13: i337.

Ning, Z., A. J. Cox, and J. C. Mullikin. 2001. SSAHA: A fast search method for large DNA databases. *Genome Research* 11, no. 10: 1725.

O'Neil, S. T., J. D. K. Dzurisin, R. D. Carmichael, N. F. Lobo, S. J. Emrich, and J. J. Hellmann. 2010. Population-level transcriptome sequencing of nonmodel organisms Erynnis propertius and Papilio zelicaon. *BMC Genomics* 11, no. 1: 310.

Papanicolaou, A, and David G Heckel. 2010. The GMOD Drupal Bioinformatic Server Framework.

Bioinformatics (Oxford, England) (October). doi:10.1093/bioinformatics/btq599. http://www.ncbi.nlm.nih.gov/pubmed/20971988.

Papanicolaou, Alexie, Steffi Gebauer-Jung, Mark L Blaxter, W Owen McMillan, and Chris D Jiggins. 2008. ButterflyBase: a platform for lepidopteran genomics. *Nucleic Acids Research* 36, no. Database issue (January): D582-7. doi:10.1093/nar/gkm853.

Papanicolaou, Alexie, Remo Stierli, Richard H Ffrench-Constant, and David G Heckel. 2009. Next generation transcriptomes for next generation genomes using est2assembly. *BMC Bioinformatics* 10, no. 1 (January): 447. doi:10.1186/1471-2105-10-447.

Parkinson, J., A. ANTHONY, J. Wasmuth, R. Schmid, A. Hedley, and M. Blaxter. 2004. PartiGene - constructing partial genomes. *Bioinformatics* 20, no. 9 (June): 1398-1404.

Pauchet, Yannick, Paul Wilkinson, Manuella van Munster, Sylvie Augustin, David Pauron, and Richard H Ffrench-Constant. 2009. Pyrosequencing of the midgut transcriptome of the poplar leaf beetle Chrysomela tremulae reveals new gene families in Coleoptera. *Insect Biochemistry and Molecular Biology* 39, no. 5-6: 403-13. doi:10.1016/j.ibmb.2009.04.001.

Schmid, R., and M. L Blaxter. 2008. annot8r: rapid assignment of GO, EC and KEGG annotations. *BMC Bioinformatics* 9, no. 1: 180.

Skinner, Me, Av Uzilov, Ld Stein, and Cj Mungall. 2009. JBrowse: A next-generation genome browser. *Genome Research* 19, no. 9: 1630.

Stajich, J. E., D. Block, K. Boulez, S. E. Brenner, S. A. Chervitz, C. Dagdigian, G. Fuellen, J. G. Gilbert, I. Korf, and H. Lapp. 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome Research* 12, no. 10: 1611 - 1618.

Stein, L. D., C. Mungall, S. Q. Shu, M. Caudy, M. Mangone, A. Day, E. Nickerson, J. E. Stajich, T. W. Harris, and A. Arva. 2002. The generic genome browser: a building block for a model organism system database. *Genome Research* 12, no. 10: 1599.

Suzek, B. E., H. Huang, P. McGarvey, R. Mazumder, and C. H. Wu. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 23, no. 10: 1282.

Terenius O, Papanicolaou A, Garbutt J.S,Eleftherianos I, Huvenne H, Sriramana K, Albrechtsen M, et al. 2010. RNA interference in Lepidoptera: an overview of successful and unsuccessful studies and implications for experimental design. *Journal of Insect Physiology in press* doi:10.1016/j.jinsphys.2010.11.006.

Wang, Z., M. Gerstein, and M. Snyder. 2008. RNA-Seq: a revolutionary tool for transcriptomics.

Nature Reviews Genetics 10, no. 1: 57-63.

Wasmuth, J., and M. L. Blaxter. 2004. prot4EST: Translating Expressed Sequence Tags from neglected genomes. *BMC Bioinformatics* 5, no. 1: 187.

Wilson, R. J, J. L Goodman, and V. B Strelets. 2007. FlyBase: integration and improvements to query tools. *Nucleic Acids Research* 36: D588-93.

Zdobnov, E. M., and R. Apweiler. 2001. InterProScan-an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17, no. 9: 847.

Chapter 6 - Analytical transcriptomic methods: case studies in non-model species

Presented herein is a collection of unpublished case-studies. Sections could be integrated in publications by collaboratiors. All investigations used transcriptomes and InsectaCentral (the web interface and the database back-end) for a variety of studies including i) producing the most data rich Insecta-wide single-copy gene phylogeny to date, ii) generating high-quality Single Nucleotide Polymorphism markers for use with the Illumina GoldenGate technology, iii) using RNA sequencing for measuring and comparing expression levels in relation to nicotine detoxification, iv) investigating genes linked with colour pattern development in *Papilio dardanus* and *P. glaucus*.

Building and utilizing transcriptomes

A number of applications in genomics require the use of a reference sequence, either to inform design (e.g. methods in molecular biology such as PCR or cloning, see Sambrook, Fritsch, and Maniatis 1989), generate probes (eg. microarray platforms DeRisi et al. 1996) or for the analysis of the data generated by the application (digital transcriptomics, reviewed in Wang, Gerstein, and Snyder 2008). It is commonplace to consider the generation of a Whole Genome Shotgun (WGS) sequence as the optimal means to drive such applications; an increasing number of groups are thus producing draft genome sequence. As genome projects from representative positions in the evolutionary tree approach completion, we can utilize comparative approaches to probe not only evolution but also function of genomic elements and genome organization (Hahn, Mira V. Han, and Sang-Gook Han 2007; Stark et al. 2007). It is still a contentious issue, however, of how representative is a specific species for the taxon it aims to represent (Wolf, Rogozin, and Koonin 2004; A.G. Clark et al. 2007). For example, the genome of the parasitic wasp, Nasonia vitripennis, was recently released (The Nasonia Genome Working Group et al. 2010). One companion paper investigated how certain detoxification families have evolved, expanded and/or contracted within the Insect phylum (Oakeshott et al. 2010). In each case, a limited number of species is available for representing each order and often a single species for a family. Even though the current data are useful in hypothesis generation and estimating, for example, diversity of glutathione-S-transferase (GST) proteins between insect orders, the variability within each taxonomic family is unknown and therefore we cannot determine whether the differences are due to life-history or simply a sampling artefact. We will also need more diverse sampling within taxonomic families (including those without a reference genome) in order to concretely make predictions of gene birth, diversification and death (Vieira, Sánchez-Gracia, and Rozas 2007). In an ideal world, we should use complete genome assemblies with each coding region verified and annotated, but almost no higher eukaryotic genome project aims for completeness (pers. observations). In reality, considering the resources needed, even for draft genomes with Next Generation Sequencing (NGS), the WGS strategy is not sustainable for the number of species needed to drive comprehensive comparative phylogenomics. Focusing on a single subtaxon, such as the 12 Drosophila Genome project with extensive genome resequencing, is one possible solution. In general, resequencing projects - where shorter and cheaper sequences are generated and assembled to a reference, well-assembled, genome of a closely related species - can be of value to the wider community. It has yet to be considered but one should, however, investigate how robust is such a procedure of cross-species/strain mapping in relation to genetic distance.

Table 1 - Abbreviations used

Full name	Abbreviation
Base pair	b.p.
Bacillus thuringiensis	Bt
copy- or complementary-DNA	cDNA
Cytochrome P450	cyp-450
Enzyme Class	EC
Evolutionary & Ecological Functional Genomics	EEFG
Generalized Linear Model	GLM
Gene Ontology	GO
Glutathione-S-Transferase	GST
International Union of Pure and Applied Chemistry	IUPAC
Kyoto Encyclopaedia of Genes and Genomes	KEGG
multi-dimensional scaling	MDS
mitochondrial DNA	miDNA
National Center for Biotechnology Information	NCBI
Next Generation Sequencing (technology)	NGS
Open Reading Frame	ORF
polymerase chain reaction	PCR
Real Time Quantitative PCR	qPCR or RT-qPCR
Quantitative Trait Locus/loci	QTL
ribosomal DNA	rDNA
Ribosomal Protein	RP
Reverse Transcription PCR	RT-PCR
Single Nucleotide Polymorphism	SNP
Short Read Archive	SRA
Simple Sequence Repeat	SSR
Whole-Genome-Shotgun	WGS

In most genome projects, the identification of coding sequence (i.e. the transcriptome and proteome) is a primary target but a secondary step. If our aim is to drive downstream applications then the approach of sequencing an entire genome, despite the far-sighted benefits it provides, may not be the necessary step in order to support a particular application. Researchers have, therefore, used partial sequencing of the genome, and in particular the coding fraction, by shotgun sequencing cDNA molecules generated from mRNA. Often the cDNA libraries are generated from mRNA of a specific tissue, developmental stage, sex and/or other biological variables. These cDNA libraries can be derived from one or multiple individuals and - along with the degree of inbreeding - this determines the number of haplotypes (i.e. chromosomes or alleles) which enter the sequencing

panel. In previous chapters, I showed how transcriptomic data could be built into reference datasets using public and custom built software. This chapter presents how cDNA-driven approaches, bioinformatic software and expertise presented in this thesis can be used to assist with addressing specific biological questions. The goal is not to investigate biological questions in depth, but to develop the approach & methodology within 'case-studies' so that a full-fledged study may be conducted in the future by myself or colleagues. As case studies, I am presenting the transcriptome assembly of a number of non-model Insect species, most of which had no genomic tools at the start of this thesis. Each of these case-studies aim, primarily, to show the wider impact of the work presented elsewhere in this thesis. On a second level, they inspired and drove a component of the bioinformatic software presented in this thesis. Briefly, these studies were as follows:

a) Towards an Insecta-wide Ribosomal protein phylogeny

As the Insecta is a most diverse taxon, making up over 80% of all of the animal kingdom (Samways 1993), the phylogenetic relationships of different orders have never been fully elucidated. The most thorough investigation has used morphological markers (e.g. “Phylogeny of Insects” in N. F. Johnson and Triplehorn 2004), which are particularly prone to the effects of homoplasy. Molecular systematics has however been hampered by narrow sampling and a deficit of markers. This case-study used ribosomal proteins (RP) sequences, a by-product of cDNA sequencing projects, to investigate phylogenetic relationships between insect orders and within certain families.

b) Gryllus campestris Single Nucleotide Polymorphism markers

The european field cricket, Gryllus campestris, (Orthoptera:Gryllidae) has promising potential for behavioural QTL mapping projects (Honegger 1981; Bretman, Wedell, and T. Tregenza 2004; Simmons 2004; Gray 2005; Jang and Gerhardt 2006) but with no available sequence data. A cDNA library from an outbred population was generated in order to first build a reference transcriptome and then mine for high-quality Single Nucleotide Polymorphism (SNP) markers suitable for use with the new Illumina bead-station genotyping technology. The resulting SNP markers are being used to establish the pedigree of a field colony used in a long-term behavioural study.

c) Manduca sexta and nicotine detoxification

Manduca sexta, a moth (Lepidoptera: Sphingidae) which has evolved a nicotine detoxification ability, has a published transcriptome which has been reanalysed for this thesis (Pauchet, Wilkinson, Vogel, et al. 2009). It is subsequently investigated in the context of altered gene expression when

challenged with nicotine. Identified candidates included a number of cytochrome P450 (cyp-450) genes, including one previously published.

d) Papilio glaucus and P. dardanus: colour pattern candidate loci

Papilio dardanus and Papilio glaucus are two butterflies (Lepidoptera: Papilionidae) which exhibit female-specific mimetic wing patterns (R. Clark et al. 2008). Reference transcriptomes for both species using wing disc cDNA libraries were generated, annotated, analyzed and used for annotation of genomic loci linked to the melanic wing patterning. Deep cDNA sequencing of a single tissue has allowed for an excellent dataset for creating a reference transcriptome for the developing wing disc. An annotated melanogenesis pathway from known genes could be constructed based on this dataset. Further, an alternative deep-SAGE strategy was performed on P. dardanus to predict candidates genes relating to the melanic pattern formation and sex-biased genes. Due to deficiencies in the experimental design a robust analysis cannot be performed but improvements on design are discussed and the same analysis pipeline can be re-run on the newly-obtained data.

Insect-wide phylogeny using Ribosomal Proteins

Introduction

Modern phylogenies can be constructed using molecular sequence data, morphological characteristics, or a combination of both. Morphological (or character) phylogenies are based on visually accessible traits, and are assessed by presence or absence of a character trait. Phylogenies that use morphological data are, therefore, particularly prone to homoplasy (similarity in trait values by chance convergence rather than shared ancestry) because of the limited number of characters used. Phylogenetic relationships can be also determined with molecular DNA-based techniques . Even though character states derived from DNA or amino acid sequences are also prone to homoplasy and need to be corrected using an evolutionary model, the larger number of such characters available should allow for a more reliable estimation of the phylogenetic relationships by increasing the signal to noise ratio. Surprisingly, the phylogeny of the major Orders of insects is still incompletely resolved. To date, most insect phylogenies have used a combination of morphological and sequence data (Whiting et al. 1997; Whiting 2002; W. C. Wheeler, Cartwright, and Hayashi 1993), although efforts to create a phylogeny of the Insecta class using molecular sequence data alone have been made (Wiegmann et al. 2009; Longhorn, Pohl, and A. P Vogler 2010; Savard et al. 2006). The success of these previous attempts has been varied, with some being restricted by

sample size or the use of genes reaching saturation too rapidly to allow for a robust evolutionary model (e.g. mitochondrial DNA in (Bae et al. 2004)). Balancing saturation and information is important for constructing robust deep phylogenies (Castresana 2000). Ribosomal RNA (e.g. 16S or 18S) has often been used in the phylogenetic construction of many taxa, including insects (Kjer 2004). They carry, however, insufficient phylogenetic information to resolve the basal relationships amongst some taxa such as Hexapoda (Misof et al. 2007). Often studies are complemented with one or more nuclear markers. The use of a single marker represents only a tiny fraction of the genome and therefore does not account for any bias associated with different rates of evolution, codon usage bias/composition heterogeneity, horizontal gene transfer (in prokaryotes at least), recombination or mutation hotspots. Only a multi-locus approach can normalize the variables and account for these biases. Likewise, contradictions seen between single-gene phylogenies may be attributable to the stochastic errors which appear due to the limited amount of information available in short sequences, such as those from single genes (Philippe and Telford 2006). A multi-locus phylogeny of Arthropoda exists, using complementary DNA (cDNA) of single-copy nuclear protein coding genes (Regier et al. 2009), genomes (Savard et al. 2006), PCR fragments of specific loci (Wiegmann et al. 2009), or RPs from EST projects (Longhorn, Pohl, and A. P Vogler 2010) but are all largely inconclusive when addressing relationships between insect families, as few species from each order are studied. It would be important therefore to develop a concrete methodology for building phylogenetic trees from EST data as new species are being sequenced. Phylogenetic studies are often directed with a particular set of standard conserved or degenerate markers, but one could consider utilizing public EST data to create a suitable dataset for phylogenetics.

To begin such an endeavour, a robust gene set must first be generated. In prokaryotes, the ribosome is comprised of more than 50 proteins, tightly bound together which make up the large and small ribosomal subunits. Eukaryotes retain an orthologous structure in the form of the (nuclear-encoded) mitochondrially localized ribosome, while also possessing the cytosolically-localized ribosome with over 70 RPs. These ribosomal complexes are involved in one of the most evolutionary conserved processes in all of biology: transcription. Such a housekeeping function provides sufficient selective constraints on the amino acid level to allow for straightforward orthology identification (Zhang and W. H. Li 2004). For this study I decided to use cytosolic RPs to complement other conserved markers used for phylogenetic studies, namely the elongation factor 1a (ef1a) and dopa decarboxylase (dopa) genes. Like ef1a, RPs are conserved to allow for design of conserved or degenerate PCR primers. Unlike ef1a, however, with several known examples of pseudogenes (NCBI web citation: http://tinyurl.com/ncbi-ef1a), there are no known RP pseudogenes in insects

(in contrast to humans). Further, unlike other housekeeping genes such as those on the upper levels of a developmental pathway cascade (e.g. wingless and dopa), all cytosolic RPs are expressed in high levels across most tissues (unlike the mitochondrial RPs) and a cadre of candidates can be acquired even from shallow transcriptome sequencing.

Methods

Raw data and reference transcriptome

Public cDNA sequences were derived from two databases provided by NCBI: dbEST (Sanger sequencing technology) and the Short Read Archive (Next-Gen sequencing technologies). Assemblies were either acquired from published work (Papanicolaou, Stierli, and others 2009) or collaborators and (re)constructed using the est2assembly pipeline version 1.03. In the latter case, an est2assembly script (trim_assembly.pl) provided a non-redundancy procedure which decreases the number of overlapping contigs and includes only those singletons that have a similarity to a known protein. Annotations on the assembled data met the specifications provided by InsectaCentral (previous chapter of this thesis). Additional beetle datasets used for RP ORF construction were generated by Dr Pauchet (U. of Exeter) with the method described in (Pauchet, Wilkinson, van Munster, et al. 2009). The Sitophilus oryzae zeamais samples were derived from adult midgut; the Callosobruchus maculatus samples from whole larvae. Additional butterfly datasets were provided by Dr Fukova (U. of Exeter) as described in the Papilio case-study. The Erynnis propertius skipper and the Papilio zelicaon butterfly datasets were provided by Scott O'Neil as per (O'Neil et al. 2010).

Curation of reference transcriptome

Manual curation on selected contigs generated by est2assembly was necessary. The est2assembly pipeline is dependent on assemblers initially created for bacterial genomic projects and only subsequently modified for eukaryotic (model) species and then further modified for transcriptomics. As a result, heterozygosity as well as alternative splicing can generate multiple partially overlapping contigs for the same protein. For manual curation, a reference protein was retrieved from one or more model species. In most cases, a single protein from Drosophila melanogaster or Bombyx mori sufficed to identify homologous contigs of the target species using reciprocal BLAST. For single copy genes, the full Open Reading Frame (ORF) was reconstructed by merging contigs following a global alignment via ClustalW as implemented in Geneious. Where two alternative proteins were possible (two different sets of contigs), both ORFs were generated since one may be the product of

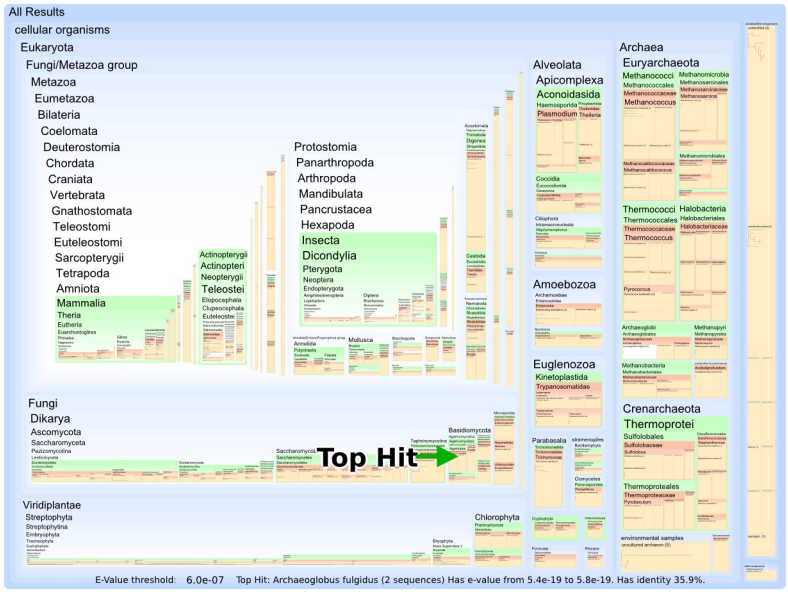


Figure 1*: Phylogeny based BLASTx search using a RpL5 protein found in T. vaporariorum. Taxon names are retrieved from a BLASTx search versus UniRef100 (with an equivalent e-value cut-off of 6e-7) and grouped according to the NCBI phylogeny. The top hit for this search is a fungal protein.*

a contaminating organism (e.g. Figure 1). Where isoforms were known to exist, dotplots (via the dottup program from the EMBOSS package Rice, Longden, and Bleasby 2000) and percent identity identified the correct ORF for the target polypeptide. Once a gene was annotated from all desired target species, a ClustalW multiple sequence alignment was generated using the protein from the fungus Neurospora crassa as the outgroup. Neighbor-Joining trees were generated to check for contaminants as they would cluster with the fungal sequence. As an outgroup for the final phylogeny, the water flea Daphnia pulex was used. Due to the difficulty of acquiring full ORFs from wfleabase (http://wfleabase.org), I constructed my own ORFs for the above genes using the est2assembly software and manual curation like the other taxa.

Phylogenetic reconstruction

A number of RPs were identified from 32 species from the Orders of Coleoptera, Diptera, Lepidoptera, Hemiptera, Hymenoptera and Orthoptera and supplemented data from 2 other genes often used for phylogenies: ef1a (another component of the protein biosynthesis machinery) and dopa. Concatenated alignments were generated using ClustalW and the translation driven alignment option (via the Geneious software). Each alignment was manually edited, especially to account for misalignments commonly found near gaps and at the N' and C' termini. Aligned regions for which the evolutionary relationship was unclear were filtered out using GBLOCKS (Castresana 2000) and

a first neighboor-joining tree gave an overview of a possible phylogenetic relationship including any potential contaminants (they would cluster with the outgroup - e.g. fungi – either via long branch attraction or because it was fungal derived). After any correction, the alignment was regenerated but replacing the fungus with D. pulex as the outgroup. Subsequently, the most likely evolutionary model fitting a supermatrix derived from concatenating all the alignments was investigated under Maximum Likelihood (ML) using RAxML under the Generalized Time Reversible (GTR) model with an estimation of the gamma parameter and a total of 10 separate inferences before choosing the one with the best likelihood value (Stamatakis, Ott, and Ludwig 2005). Each gene in the supermatrix formed a separate partition allowing parameters to be estimated independently. The waterflea, D. pulex, was used as the outgroup.

The rapid of option of RAxML (-f a) with 100 bootstraps was first used to test for best ML values and highest support. It would, however, often report very low bootstrap values due to a non-optimal ML tree being used. As a single tree inference is performed in the rapid option, I used a manual method for producing bifurcating trees (-f d with 10 inferences; -f d with bootstrap -b; -f b to reconsile bootstrap and final trees). The M. cinxia data showed a high proportion of missing data. Previous authors (Savard et al. 2006) showed that composition bias (i.e. composition heterogeneity) in third codon sites provided misleading results. Therefore, in the final alignment, RY-coding on third codon positions and removal of M. cinxia was used to improve initial trees. The RAxML program using performed 1,000 bootstraps and found the most optimal ML bifurcating tree with bootstrap support value. For both the bootstraps and the final ML tree, 10 different starting trees were used to initiate the ML landscape search. For comparison, a method utilizing ProtTest/jModelTest model selection (Abascal, Zardoya, and Posada 2005) and the PhyML software (Guindon et al. 2009) was performed but the RAxML was chosen for the ability to consider a single supermatrix with partitions, utilize outgroup sequences and the observation that the GTR model is the most robust across different datasets (A. Stamatakis TU-Munich, pers. communication). For the composition heterogeneity tests I used matched pairs of symmetry as proposed by (Bowker 1948). It was implemented in custom software coded in C as kindly provided by Dr. Lars Jermiin and utilized a methodology elaborated in (L.S. Jermiin et al. 2008). I subsequently altered the code to allow for investigating the effects of recoding per-se, recoding/removing specific codon sites, including/excluding taxa/genes and also a sliding window approach. A custom written script, rycoding.pl, was used to alter the alignment and investigate effect of criteria such as implementing i) RY-coding, ii) removing specific taxa, iii) removing all positions with gaps (rather than those present in > 50 % of the sequences) and iv) removing the 3rd codon positions.

Table 2 List of Ribosomal Proteins in the dataset and their ORF length

Order	Coleoptera								Diptera				Heteroptera		Hymenoptera			
Marker / Species	CM	CT	GV	SO	TC	AA	AD	AG	DM	DS	DY	LL	AP	TV	AM	LT	NG	NL
L21	477	477	477	477	477	483	480	477	477	477	477	477	474	483	477	477	477	477
L3	1,245	1,242	1,239	1,236	1,227	1,245	NA	1,251	1,248	1,248	615**	1,260	1,224	1,242	1,245	1,236	1,257	NA
L31	372	372	372	372	372	372	372	372	372	372	372	372	372	375	369	369	372	372
L5	894	909	891	888	900	891	NA	891	897	897	897	891	900	888	894	891	891	852
S11	456	453	453	465	456	456	456	456	465	465	435*	456	456	447	465	465	462	NA
S12	420	417	417	420	420	420	411	411	417	417	417	420	408	411	423	420	420	399
S13	453	453	453	453	453	453	NA	453	453	453	453	351*	453	489	435*	426*	453	429*
S14	453	453	453	453	453	453	456	456	453	453	453	453	453	453	453	453	453	NA
S16	453	459	450	456	444	438	411*	444	444	444	444	447	423*	438*	444	444	444	393*
S17	NA	393	393	393	393	390	393	393	393	393	393	390	390	387	393	390	393	393
S18	456	456	456	456	456	456	456	456	456	456	456	456	456	456	456	456	456	456
ef1a	1,386	1,386	1,386	1,389	1,386	1,389	1,389	1,389	1,386	1,386	1,386	1,389	1,386	1,386	1,383	1,386	1,383	NA
Total bp	7,065	7,470	7,440	7,458	7,437	7,446	4,824	7,449	7,461	7,461	6,183	7,362	7,395	7,455	7,437	7,413	7,461	3,771

Order	Lepidoptera												Orthoptera		Outgroup	Aligned	
Marker / Species	BA	BM	EA	EP	HE	HM	HN	MC	MS	PD	PG	PZ	GB	GC	DP	33	species
L21	477	477	477	477	477	477	477	441*	477	477	477	477	477	477	480	483	b.p.
L3	1,236	1,239	1,236	1,254	1,236	1,236	1,236	1,236	1,242	1,239	1,200	1,236	870**	1,215	1,236	1,260	b.p.
L31	372	372	372	360	372	372	372	195**	372	372	372	372	372	372	372	375	b.p.
L5	891	897	891	898	888	888	888	891	897	894	897	564**	624**	897	906	909	b.p.
S11	456	456	456	456	456	456	456	364	456	456	456	456	456	456	447*	465	b.p.
S12	417	417	417	264*	417	417	417	273*	417	417	417	417	423	423	414	423	b.p.
S13	453	453	453	453	453	453	453	330*	453	453	453	453	465	465	453	489	b.p.
S14	453	453	453	423*	453	453	453	378*	453	453	453	453	453	453	453	456	b.p.
S16	453	453	453	NA	453	453	453	408*	453	447	453	453	447	432*	462	462	b.p.
S17	399	399	399	399	399	399	399	312*	399	399	399	399	393	393	387	399	b.p.
S18	456	456	456	456	456	456	456	456	456	456	456	456	456	456	453*	456	b.p.
ef1a	1,389	1,389	1,389	1,389	1,389	1,389	1,389	1239*	1,389	1,389	1,389	1,389	1,386	1,386	1,389	1,389	b.p.
Total bp	7,452	7,461	7,452	6,829	7,449	7,449	7,449	5,920	7,464	7,452	7,422	6,561	5,328	7,425	7,452	7,083	b.p.

* Complete ORF not determined but included in the analysis

** Complete ORF not determined and object not included in the analysis

Results

Manual curation

From generating a single gene (ef1a) phylogeny I determined that including taxa with less than 50 % of the ORF produced topologies which would violate the known monophyly of most insect orders. The effect was the same for alignment partition less than 1 kb (after removing uninformative sites), meaning that the GTR Markov model was not able to parameterize the model with such small datasets. This was assumed to be exaggerated due to compositional heterogeneity but no correction were done at that early stage. Unlike some genome-wide studies which rely on automated alignments, I chose to reduce the level of noise and be confident on the nucleotide sequence of the genes. I excluded, therefore, all samples which had less than 75 % of the ORF. Corrections via manual curation of each protein were often needed. The annotation approach begun with GenBank mRNA data and was checked with InsectaCentral data. The first curation pass was accomplished by an undergraduate student (see contributions) but due to missing taxa and an error in the Heliconius clade, I re-annotated all ORFs. Sequences were often partial but by searching on InsectaCentral I was often able to extend the ORF. Further, even for some species with a sequenced genome (Aedes and Apis sp), genes such as the Elongation factor could be better characterized from the InsectaCentral data rather than gene models automatically predicted and available in NCBI. On the other hand, in an ef1a gene from InsectaCentral, there was one case of a chimeric read derived from Sanger data. The 3' end of IC7460AaEcon4 was chimeric and according to the coverage graph provided by InsectaCentral it was supported by only one read. This probably accounted for the alternate contig IC7460AaEcon1186, which partially overlapped (the overlap on IC7460AaEcon1186 also supported by single read), provided the missing end. The approach of using the coverage graph was useful for determining which sequence to use in other cases of partially overlapping contigs. When using GenBank mRNA data, the most common error that was encountered was sequencing errors associated with Sanger reads (e.g. ef1a from Bicyclus anynana). Because InsectaCentral's contigs are derived from a clustering of the data available on GenBank, it was possible to correct them. For the RPs, I relied mostly on InsectaCentral data unless a full-length mRNA had been manually annotated and submitted to GenBank (e.g. Drosophila melanogaster, Bombyx mori). Even though all genes were housekeeping genes and overall conserved, some alignments had unconserved N- or C-terminal ends (note that all alignment inspections were done with the ORF and the translation in mind), for example the RpS16 N-terminus end. For that reason, the Gblocks approach would be later utilized.

Once a putative alignment was available, two important types errors can still exist. From my experience, "self-correcting" frameshifts are not uncommon in Sanger-derived EST contigs. In this case a frameshift can, by chance, be corrected by another frameshift a few amino acids downstream. Because of coverage, NGS-containing datasets do not exhibit this effect. By looking at the translation of all forward frames in a multi-species alignment it is possible, however, to detect such cases and correct them. Even though the program prot4EST performs this automatically, it fails to amend self-correcting frameshifts. Such cases were commonplace in species with low EST coverage such as Anopheles darlingi, Melitae cinxia or Nasonia giraulti. The second error is best exemplified with the dopa decarboxylase gene, a member of a gene family. A number of sequences were potential orthologues but after generating a protein family tree of all potential paralogues from the model species it was possible to determine that a number of them were indeed paralogues but clustered due to overall sequence similarity. There was an insufficient number of true dopa decarboxylase sequences to add any value to an insect phylogeny. In total, 12 markers from 33 taxa were identified, with a total of 7,053 aligned base pairs (Table 2).

Initial tree generation and compositional heterogeneity

First, a single ML tree with 32 species was generated (Figure 2A). Each taxonomic order was monophyletic and sister to the other orders. Three nodes/arrangements were poorly supported, all at the family level: i) the first branching of Diptera between the Drosophila sp., the sandflies and the mosquitoes; ii) the second branching of Lepidoptera and specifically the arrangement between moths and true butterflies after the branching of skippers; iii) the coleopteran branching between Tribolium castaneum and chrysomelids. This tree, however, was affected by compositional heterogeneity; i.e. different utilization of amino acids due to different biases in nucleotide compositions in different groups (Savard et al. 2006): the Hymenoptera did not show as the basal holometabolous order and non-holometabolous insects would cluster with holometabolous insects. RY coding of 3rd codon positions, as per other similar studies such Wiegmann et al and Longhorn et al (Wiegmann et al. 2009; Longhorn, Pohl, and A. P Vogler 2010), provided a more realistic branching order (Figure 2B) but a) compositional heterogeneity may still exist even with excellent bootstrap support values; b) some deep branching bootstrap support values were not as high as wished for. Single gene tree investigations showed that partitions less than 1 kb were poorly supported and giving a false topology (nb alignment sizes included all and only informative sites but some branches would have fewer informative sites than the total available).

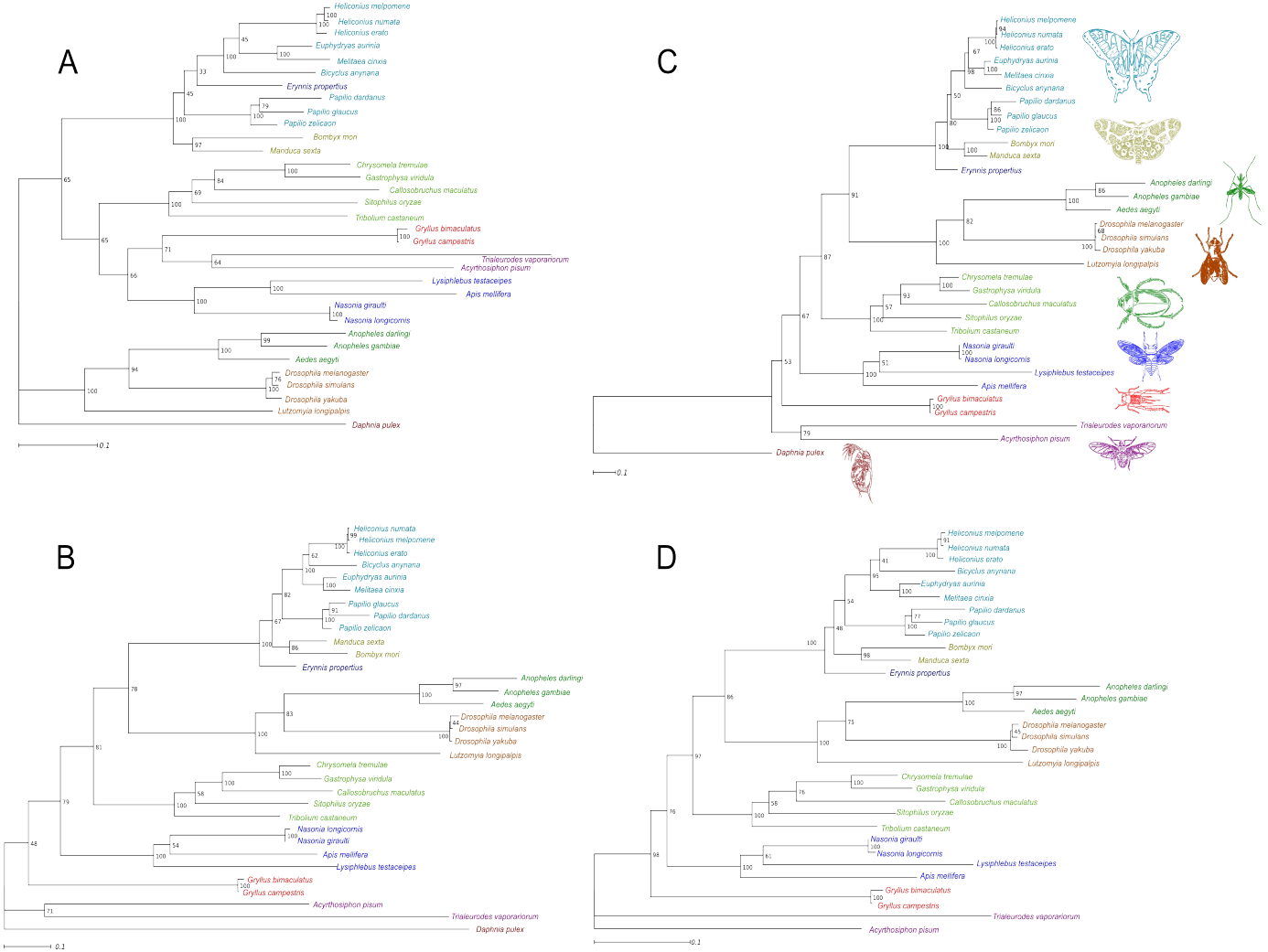


***Figure 2**: Maximum Likelihood phylogenetic trees using RAxML (A-C; rooted) and PhyML (D; not rooted) with bootstrap support values. A) Tree generated using full alignment without correcting for composition heterogeneity. B) Same data as (A) but with 3rd codon site recoded with R/Y IUPAC codes. C) Same data as (B) but 1st codon site removed. D) Same data as (C) but without the outgroup and processed using PhyML to produce an unrooted tree.*

Further, a 3rd party analysis of the data presented Wiegmann et al (Wiegmann et al. 2009; Jermiin, CSIRO, pers. communication) showed that RY coding was not sufficient to alleviate compositional heterogeneity and the number of informative sites was very low even though published support values were high. This can be due to a poorly trained Markov model still being a good fit to a small dataset because few alternative hypotheses are possible.

Compositional heterogeneity

Table 3 – Investigation of compositional heterogeneity

Codon site \ % of pairwise comparisons significant at p<0.05	Original alignment	Without Diptera	After RY coding	After RY coding and removing Diptera
All codon sites:	92.60%	89.20%	57.60%	47.10%
1st codon sites:	59.50%	37.50%	40.70%	12.60%
2nd codon sites:	5.90%	5.50%	5.90%	6.50%
3rd codon sites:	92.80%	90.20%	54.00%	47.40%
1st & 2nd codon sit	56.80%	34.50%	41.10%	12.90%
1st & 3rd codon sit	92.20%	89.50%	57.00%	46.80%

Compositional heterogeneity was then extensively investigated. The results showed that across the entire alignment extensive compositional heterogeneity exists in 1st codon sites as well as 3rd codon sites (in 59.5 % and 92.8 % of the sites respectively, Table 3). In such cases there are two options: exclude the affected sites or recode them. Further, the effect from specific taxa can be detected. Mild compositional heterogeneity on the second codon site was identified as miscuration of the Tribolium castaneum RpL3 gene (the longest RP in the dataset). After correction, heterogeneity for 1st and 3rd codon sites still existed. Bias for the third codon position was spread in the entire phylogeny. The 1st codon site was most significant for the Drosophila clade but after RY coding (when synonymous) and removing the dipteran clade, a significant proportion of the pairwise-comparisons still violated GTR assumptions. Other recodings were investigated using a custom function in rycoding.pl. All other types of coding (AB, ACK, AGY, ATS, CD, CGW, CTR, GH, GTM, KM, SW, TV) were investigated but were less effective than RY.

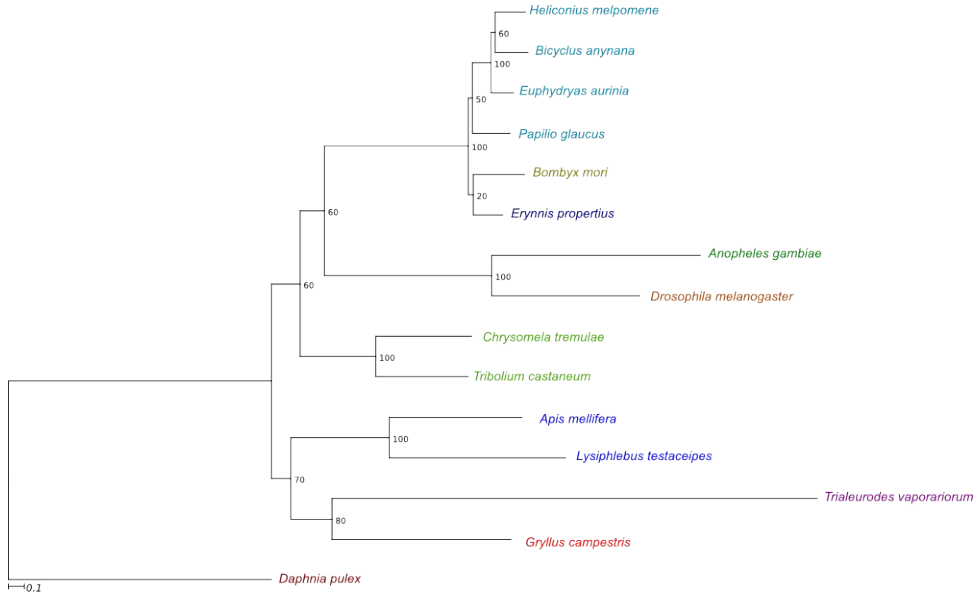


***Figure 3**: Maximum Likelihood phylogenetic tree using RAxML with D. pulex set as the outgroup. Dataset was same as Figure 2B but with fewer taxa. Node labels are bootstrap support values. Note the higher support values and the grouping of the hemi-metabolous insects with the Hymenopteran clade.*

Investigation of compositional bias

Table 4 – Likelihood search

Method	Likelihood score	Sites
No change	-103599.11	7,263
3rd codon recoding; no partitioning	-61523.70	4,842
3rd codon recoding; codon partitioning	-59495.06	4,842
1st & 3rd codons recoded with RY. 1st only if synonymous change	-57864.61	4,842
1st & 3rd codons recoded with RY	-47834.51	4,842
1st codon removed, 3rd recoded; codon partitioning	-38506.69	4,842
2nd codon only; no partitioning	-13147.27	2,421

To complement the above findings, likelihood values where used to test how the generated model fitted the data after transformation (Table 4). They were compared from trees constructed from the various composition heterogeneity datasets and by also considering the partitioning variable: i) no

partition, ii) partition by codon iii) partition all short Rps (<1,000 bp) together and the other genes separately. Trees using all possible combinations yielded a range of likelihood values with the best one being the one with 1st codon positions removed, 3rd codon positions RY coded and partitioned using codon information (i.e. all genes belonging to the same partition; Figure 2C). This likelihood was significantly higher than RY coding the 1st codon positions but the would not be directly comparable as there are 50 % more sites than in the stripped alignment. However, alignments of equal length can be compared and for example using RY coding for the 3rd codon only and using no partitioning. Even this value compared favourably with the initial non-recoded, non-stripped alignment. I considered that the clustering of the RPs and ef1a as a single unit is thought to be appropriate because the rate heterogeneity between codon sites would be higher than between genes. Further the similar purifying selection forces are probably acting on these genes as they all are single-copy housekeeping genes.

In all trees where compositional bias was addressed, the topology did not differ but branch lengths were different (Figure 2B-2D). Both of the hemimetabolous and the holometabolous clades were monophyletic. The Hymenoptera clade was indeed the most basal holometabolous order. The grouping of the Mecopterida (Diptera + Lepidoptera in this study) and Neuropteroidea (Hymenoptera and Coleoptera in this study) is also preserved and hemimetabolous insects are positioned outside the holometabolous clade. It is important to note that within the lepidopteran clade, the skipper butterfly Erynnis propertius is more basal than the split of the derived moths and true butterflies but some of the discarded trees placed them after the split. Overall, the tree topology was not robust in one more node: the split between Orthopteran and Hymenopterans (and therefore the split between hemimetabolous and holometabolous insects) with a bootstrap value of 56 %. Using the same alignment as input to the phyML program (which allows other models than the GTR), removing the outgroup (phyML accepts no outgroups) and using the HKY85 model the topologies were identical but bootstrap support value for that node was 98 % (Figure 2D). Using jModelTest, however, the GTR model was the one with the best AICc value. Finally, the effect of including only a small number of taxa, as per Savard et al (Savard et al. 2006), was investigated using 2nd and recoded 3rd codon sites (Figure 3).

Discussion

This study was focused on determining i) the extent to which EST data can be used for phylogenetics of insect orders; ii) phylogenetic branching of the major holo- and some non holo-metabolous insect orders and iii) providing novel insights within orders. This study used the

nucleotide sequence of 12 conserved genes but can be extended by using other data. Indeed, use modern phylogenetic software allows for a compound strategy; even both proteins and nucleotide sequences can be considered. Deep phylogenies (e.g. between phyla or kingdoms) can utilize the protein sequence to allow for sufficient degree of phylogenetic signal but avoid the noise associated with saturated changes on the nucleotide level. Mid-depth phylogenies, such as the one presented herein, would not acquire sufficient phylogenetic signal on the protein level and will have to utilize the nucleotide sequence. For species delimitation, faster evolving genes, such as mitochondrial DNA would be used and a reconciliation approach would construct a final tree.

In deep and mid-level phylogenies, additional care must be exercised to account for i) ensuring the alignment columns are orthologous; ii) nucleotide substitutions are accounted for via an appropriate evolutionary model, preferably one robust enough to account for the diversity of organisms investigated; iii) compositional heterogeneity is investigated (Foster and Hickey 1999; L.S. Jermiin et al. 2008). This work has addressed all of these criteria. Work by other workers (Savard et al. 2006; Wiegmann et al. 2009; Longhorn, Pohl, and A. P Vogler 2010) has addressed the same phylogenetic relationships presented here and this study should be placed in this context. One of these studies (published after this one was completed), used a similar method to obtain markers: Longhorn et al (Longhorn, Pohl, and A. P Vogler 2010) used RPs to create an supermatrix of 10 Kb (compared to 7Kb in this study).

Even though InsectaCentral provided UTR information and some regions such as the 5' UTR might be useful for constructing shallow phylogenies, these were not used in this study since the ORF's third codon sites were saturated. Initial protein based trees had too few informative sites to provide a robust tree. I used, therefore, a ORF nucleotide alignment that was driven by the protein alignment. Because an evolutionary model was used, like in any model training, signal to noise ratio must be maximized in phylogenetic reconstruction since low signal to noise levels provide low confidence (as exemplified by a bootstrap test) on the resulting tree: either in the topology or the branch lengths. Removing, therefore, site groups (blocks) for which we have poor confidence reduces the noise even though some loss of information can occur (Castresana 2000; Dutilh et al. 2007).

In ML phylogenetic reconstruction, there are a number of evolutionary models to use but essentially they are special cases of the Generalized Time Reversible (GTR) model. Each of the special cases of GTR allows for faster generation of the tree and reduces the number of parameters estimated. The disadvantage, however, is that additional assumptions must be made. In this work there were sufficient informative sites to avoid overparameterization (i.e. overfitting) for most nodes and thus

the GTR + gamma model was used. The gamma parameter was estimated from the data in order to account for evolutionary rate heterogeneity between sites and, in the final tree, it was estimated separately for each codon site. The invariable sites parameter was not used: estimating the proportion of invariable sites is another method for accounting for rate heterogeneity but work has shown that it influences the gamma parameter in such a way that simultaneous parameterization of both gamma and proportion of invariable sites is impossible (Yang 2006). However, one node, the deepest, did show evidence from parameterization as the bootstrap support values were much lower with GTR than the special case of HKY85.

The topology however did change significantly when I attempted to include only a subset of species per order (Daphnia pulex as the outgroup, Chrysomela tremulae, Tribolium castaneum, Anopheles gambiae, Drosophila melanogaster, Apis mellifera, Lysiphlebus testaceipes, Bicyclus anynana, Bombyx mori, Heliconius melpomene, Erynnis propertius, Papilio glaucus, Gryllus campestris; Figure 3) with the recoded dataset: the Hymenoptera would cluster with the hemimetabolous insects with high bootstrap support (70 %) but the hemimetabolous insects would form their own monophyletic cluster. Indeed the effect of the number of species and outgroup played a significant part: if the outgroup is removed so that it mimics the data of Savard et al (Savard et al. 2006) (they used as an outgroup an Orthopteran or an aphid), the tree appears so that the holometabolous insects are monophyletic. If D. pulex, a non-insect arthropod is included, the distance between hemimetabolous and holometabolous insects is much greater.

Compositional bias was an unavoidable feature of this dataset and unlike the other aforementioned authors, I believe the issue is far more complex than a quick and dirty solution offered by recoding. The GTR model (and the all the other ML models which are special cases of the GTR), requires that the probability of e.g. A to change to a G is equal to the probabilty of G to change to an A. This is the Reversible concept of GTR. Further, a nucleotide's rate of change (Time in GTR) is constant, i.e. the probability is the same along the evolutionary history or in other words across branching events. Compositional heterogeneity (as detected by codon usage bias) can invalidate such assumptions and bootstrap tests will not show the effect as the trained model may still fit well with the data. It is unsettled exactly how changes in base frequencies within a species leading to compositional heterogeneity among species can evolve, but it may occur via a species having different DNA repair mechanisms than others (Sharp and Matassi 1994; Filipski 1987) but other processes might be at play as well (Rocha and Danchin 2002; Eyre-Walker 1996). During the protein prediction step of est2assembly, codon usage tables are constructed by aligning the ORFs from EST data to resulting proteins and it can be seen that differences exists between lineages (data

available but not shown). Even though in the future I would be interested creating a test based on codon usage tables, differences can also be detected for specific alignments using the rate heterogeneity tests conducted here. Indeed, when considering all species, codon sites 1 and 3 were significantly affected and measures were undertaken to account for this difference unlike other studies which considered only the 3rd codon site. This allowed us to estimate both the topology and the branch lengths with less violations of the GTR model. The case may be, however, that the GTR assumptions (and therefore all nested models) do not hold: evolutionary change is unlikely to be constant across time when such phylogenies are considered. Generally, however, compositional bias, or more appropriately termed compositional heterogeneity, warrants further and thorough investigations, possibly using software employing specific tests. I employed only matched-pairs tests of symmetry here but there are other tests such as those detecting internal or marginal symmetry in contingency tables (Ababneh et al. 2006; Stuart 1955) to accurately detect bias for certain genes and codons. Indeed, with the sequencing of more transcriptomes, question regarding classes of genes or species being more prone to composition heterogeneity (e.g. potentially ribosomal proteins in Lepidoptera (Landais et al. 2003), ancient genes, genes of specific function, sequence conservation or expression level) can be of wider interest.

In summary, my phylogeny with recoded 3rd codon sites and inclusion of all taxa agrees with published results of Wiegmann et al, Longhorn et al and Savard et al (Wiegmann et al. 2009; Longhorn, Pohl, and A. P Vogler 2010; Savard et al. 2006) with monophyly of each order. The removal of 1st codon sites addressed the issue of compositional heterogeneity to the effect of decreasing the number of informative sites. Other authors ignored any effect on the 1st codon sites. Further, Savard et al (Savard et al. 2006) used only 6 taxa and a whole genome approach, which would contain levels of noise affecting branch length (their study was primarily focused on topology). They also used a more closely related ougroup and it may be that their tree's topology/branch length would have changed had they used a non-insect arthropod (as in Figure 1). Wiegmann et al (Wiegmann et al. 2009) used more taxa, (30 - almost as many as this study but from more orders) but had fewer informative sites. Due to composition heterogeneity, however, both the Wiegmann study (Jermiin, pers. communication) and this study have an relatively low number of informative sites as evidenced from the bootstrap values. The bootstrap values for Wiegmann were much higher in the deeper nodes than this study. Non-parametric bootstrap is, however, not an indicator of confidence of the tree being correct: it is the level of confidence we have that the data fit the model. Over-parameterization is one possible explanation of this but Wiegmann et al (Wiegmann et al. 2009) describe how the software ModelTest was used to determine which model

gives a better likelihood estimate in relation to the number of parameters estimated from the data. As a result, they used the most parameter-rich model available, GTR with estimation of both the gamma and number of invariable sites parameters. Further, in this study, bootstrap support values in the non-recoded alignment were much higher but gave a false topology; likewise for the recoded but species poor tree. Further, Savard et al (Savard et al. 2006) used sequenced genomes (as available in 2006) and, in contrast to them, I used a more distant outgroup in order to detect the branching of Hemiptera and Orthoptera and a four-fold number of taxa. The study by Wiegmann et al (Wiegmann et al. 2009) also used a low number of species per order. Further, not all markers were successfully amplified from all species, some species such as T. castaneum had all 6 markers and other such as Strangalia bicolor were limited to 3 markers. Even though my dataset presents the same topology as theirs, unfortunately, however, my dataset does not offer the high statistical confidence that they present.

Conclusion

Having a phylogenetic framework can be essential for functional or evolutionary work but constructing one is still laborious. The RPs as phylogenetic markers have been informative; discriminating for both the order and the species level. Because of measures taken to address issues with composition heterogeneity, however, the number of informative sites has not been sufficient to produce a robust phylogeny. The RP and ef1a genes were straightforward to acquire in a number of species. One important take-home message from this case study is that a single informatics author can produce an extensive phylogeny without recourse to wet-lab experiments or use of data derived from other phylogenetic studies. Indeed, this work could and should be extended by using the data generated by Longhorn et al (Longhorn, Pohl, and A. P Vogler 2010) and other workers as well as more genes derived from EST data. What is important, however, is to use a robust methodology so that phylogenies can be produced as more data become publicly available. Indeed, the concept of a phylogeny must not be static. In the context of InsectaCentral it would be of interest to have the ability to deposit alignment and trees in a fashion that would be compatible with other similar databases such as TreeBase (Piel, Donoghue, and Sanderson 2000) and ScratchPads (V. Smith et al. 2009). The same framework would also drive gene-family phylogenies (as opposed to these taxonomic ones) and aid in orthology identification in the same manner that the dopa decarboxylase genes were annotated. This issue is further outlined in the final Discussion & Outlook chapter of this thesis.

Gryllus campestris Single Nucleotide Polymorphism markers

Introduction

Crickets have been well studied in the laboratory, revealing that they have complex forms of sexual selection whereby females choose between males according to their songs (Simmons 2004), males fight (Bretman, Rodríguez-Muñoz, and T. Tregenza 2006), females manipulate sperm from several males to favour unrelated males (Bretman, Wedell, and T. Tregenza 2004; Bretman, Newcombe, and T Tregenza 2009), females lay eggs faster when mated to dominant males and so forth. Although this provides a number of insights into the behaviour and physiology of crickets in the laboratory, we have almost no idea how important these various aspects are in the real world. Current work in the laboratory of Prof. Tregenza (U. of Exeter, UK) is utilizing 96 cameras to monitor a number of life-history traits in a small population of individually marked crickets: e.g. longevity, mating partners, male competition and mating display outcomes. Synthesis of these data in a genetic context can provide valuable insights in reproductive success of individuals and the heritability of these life-history traits. Current genetic data is based on genotyping every adult cricket in the test-site from 2006 to the present with 13 Simple Sequence Repeat (SSR; microsatellite) markers. The analysis of this particular SSR dataset, however, has a significant degree of uncertainty due to null alleles, potential homoplasy and misassignments, limiting thus the power of any association tests and heritability estimates. Therefore SNPs (single nucleotide polymorphisms) were investigated as an alternative for distinguishing individuals in the study, relating parentage and estimate fitness.

The improved parentage assignment that we can achieve with an NGS SNP study will allow us to investigate heritability of traits with much greater power than we are currently able to do. Additionally, we can use the same markers with traditional linkage mapping approaches to measure genome-wide heterozygosity and determine whether it predicts mate choice or the reproductive success of mating pairs. This will allow a direct comparison of the importance of genes relating to mating success and relatedness to reproductive success. Because of weak correlations between heterozygosity at a small number of markers and genome wide heterozygosity, this type of approach is only possible with a large number of markers. An Illumina NGS BeadStation utilizes a targeted sequencing strategey whereby a sequence-specific primer is used to specifically amplify a genomic DNA fragment containing a known SNP. This 50-mer, sequence-specific oligo primer allows one to include 96 markers which are more than sufficient for the specific application. Therefore, we designed the experiment to use half of those to infer parentage and measure relatedness between

parents with the other half. This gets around the problem that studies using the same suite of markers for inferring parentage and relatedness. This bias arises because paternity is more likely to be assigned if a male carries different alleles to a female, and it has been largely ignored in the relevant literature (Tom Tregenza, U. of Exeter, pers. communication).

Methods

Sequence data

Transcriptome libraries from the brain tissue of 20 individuals of mixed sex were generated by Dr Y. Pauchet (University of Exeter) using the same SMART/Trimmer cDNA library generation and normalization methodology as described in (Pauchet, Wilkinson, Vogel, et al. 2009). They were subsequently processed with est2assembly v.1.03.

Single Nucleotide Polymorphism markers for the Illumina Bead Station

For this dataset, SNP markers were predicted using ic_create_snps from the est2assembly package v 1.03 using highly stringent parameters. Markers were identified if only two alleles existed with the minor one supported by at least 5 reads and therefore at least 10 reads were needed for a particular alignment column to call a potential SNP. Any SNP which had a non-invariant 100 bp flanking region was ignored. The 100 bp up/downstream of the SNP were reported along with 66 bp and 33 bp segments. Also reported was whether the SNP was non-coding or part of a codon and whether it thus caused a synonymous or non-synonymous mutation. Identified SNPs and their flanking sequences were scored for design by the proprietary Illumina software from a scale of 0.0 – 1.0 with 0.8+ being acceptable. Due to sequence conservation, sequencing primers are more likely to amplify in a natural population if they were part of an ORF but only if an intron does not interrupt the sequence between the two primers. As the data were generated from cDNA and the intron/exon boundary is unknown for each marker, it is common practice for one to identify the intron/exon boundaries bioinformatically (via conservation with a sequenced genome of the same or related species) or opt for non-coding markers. As no Orthopteran genomes have yet been sequenced, I investigated the suitability of using a genome from another order. Searches in FlyBase, showed that 10 randomly picked genes showed no conservation of intron/exon boundaries between Drosophila melanogaster, Anopheles gambiae and/or Apis mellifera (e.g. http://tinyurl.com/35tpobj and http://tinyurl.com/386jp9e). It was, therefore, unlikely that transferring intron/exon boundaries from another genome would be appropriate. Further, no wet-lab capacity existed to verify those SNP

markers. The alternative approach was therefore to decrease the length available for the design of the sequencing (Illumina expects a 201 bp fragment). Aided by the design scores provided by Illumina, a range of markers were chosen with length of 66 ~ 201 bp. As a result, markers were then sorted by design score and 30 coding and 66 non-coding markers were chosen.

Results and discussion

SNP marker selection for determining reproductive success and trait heritability in a wild insect population using Next Generation Illumina technologies

Table 5 Number of conserved SNP markers in relation to allele frequency

Frequency of minor allele (%)	Number of markers
10 ~ 19	287
20 – 29	319
30 ~ 39	270
40 ~ 49	279
50	75

InsectaCentral, via the est2assembly pipeline, provides for each library a list of all high quality SNPs (Figure 4). As only 96 markers were needed for the project, a sub-selection of markers was performed. In the first round, very stringent criteria were used: I selected those which had at least 5 reads supporting the minor allele (i.e. at least 10 reads supporting the SNP) and 100 invariable bases up/downstream of the SNP gave a total of 4,354 SNPs of which 2115 were transversions and 2239 where transitions. As the cDNA library was not generated from the same individuals as the ones to be genotyped (but a random sample from the same outbreeding population), further criteria were applied to ensure that the marker would work across as many individuals as possible. As InsectaCentral has identified the SNPs belonging to coding regions, 1,375 coding markers were selected (523 causing a transversion and 852 a transition). In order to increase the chances that a SNP would be polymorphic in the genotyping panel and would amplify, the invariable region was increased to 100 bp and the markers were tabulated according to the frequency of the minor allele (Table 5). As the markers would be used for a paternity study, we would not expect balancing selection to cause any bias and therefore markers with minor allele frequency of more than 40 % were sent to the sequencing facility to assign design scores (according to an Illumina proprietary algorithm). Markers which had a design score of more than 0.8 and where from different contigs were selected for the final genotyping.

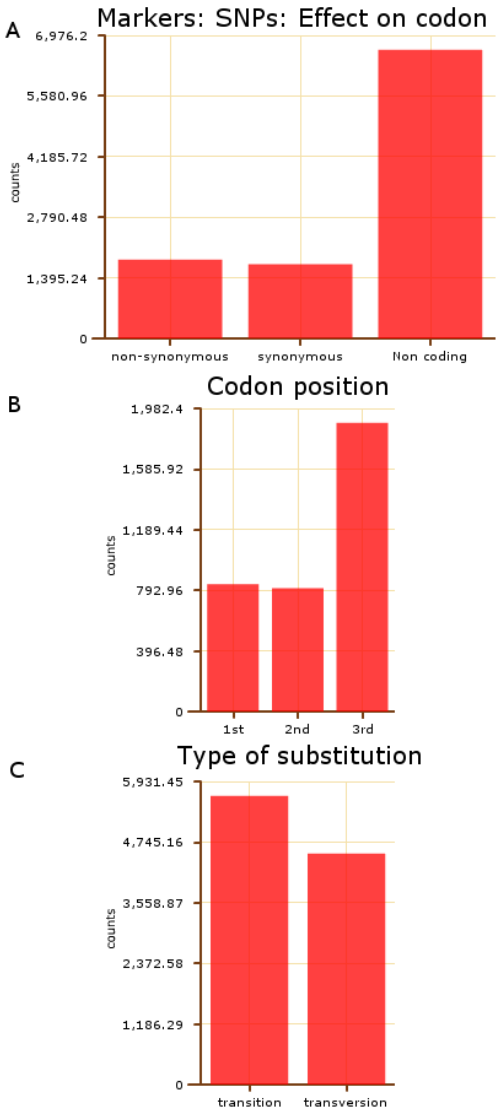


***Figure 4**: Cricket SNP markers as identified with est2assembly and automatically visualized in InsectaCentral for the G. campestris libraries. Markers can be categorized according to A) effect the SNP has on the amino acid; B) codon site; C) whether the SNP causes a transition or transversion. Of these markers a subset was chosen for the Illumina Bead Station using stringent criteria (see text).*

Expected data

The specific protocol is sufficiently novel that few studies have been published and none were found to be from insects even though studies on plants have been performed. One published study is on mapping the evolution of human-bred rice lines (T. Yamamoto et al. 2010). Via Illumina-based re-sequencing, they first produced a ca. 5.89 Gb (ca. 15.7 x fold genome coverage) sequence from a previously unsequenced rice cultivar. From that they managed to predict 67,051 candidate SNPs (rice has a very low degree of polymorphism) of which 1,917 successful markers were used for genotyping 151 rice cultivars using the Illumina Bead station.

Their method differed in a number of ways. First SNP identification for the rice study was accomplished by mapping of individual reads to a reference genome; in our study we used the

databased assemblies which had SNP predicted from the existing alignment of the reads making up each contig. Like this study, however, they used existing gene models to classify SNPs as coding and whether it was causing a (non-)synonymous change. InsectaCentral, however, already holds this information via the est2assembly pipeline. Extracting this information is possible via the web-interface even though the advanced filtering options employed here are not yet available. Third, they tolerated far lower Illumina design scores (0.4) versus our study (0.8) which may be the reason why only 1,917 markers met the fluorescence criteria after the run was complete (a 71 % success rate). It would have been assumed that the low degree of polymorphism and good quality gene model annotation would ensure that most markers would have met the fluorescence cut-offs. Like this study, they utilized genomic DNA as a genotyping substrate but they had a-priori information of intron-exon boundaries (via gene-models annotated using the genome).

SNP genotypes

Overall the resulting dataset was of poor quality. Around 75% of the SNP markers were verified by Dr Jon Slate (University of Sheffield Sequencing Facility). The call confidence rates were, however, low and indicative of low quantity of DNA template. From the available data, there was no evidence that coding versus non-coding region putative SNPs differed in how they convert to 'true' SNPs. The main limitation in analyzing the data was with the Illumina GenomeStudio software: it calls genotypes based on clusters of fluorescence values. For each SNP, all samples are analysed together, with the idea that when plotted they form three 'clusters' representing the three possible genotypes (Figure 5). In this particular dataset, many of the clusters were dispersed and overlaps made genotype calls difficult or impossible.

Conclusion

At the time of writing, the sequencing experiments to generate reliable SNP calls for the paternity study were not yet completed. Current data did show that circa 75 % of the SNP markers were valid and amplifiable (Slate, pers. comm.). The method employed here provided highly conserved SNP markers but the available sequence provided more than sufficient candidate markers. It is likely that many of these markers are not neutral but for an application on a pedigree it will not be of importance, since the identification method relies in excluding potential parents by SNP presence/absence, not frequencies of SNPs. Indeed balancing selection might assist in maintaining high levels of polymorphism and therefore chosen markers had high minor allele frequencies. For designing such conserved, potentially not neutral markers, an alternative, less expensive, strategy

which could be considered is the sequencing of a single lane of Illumina, de-novo assembly and a BLASTx-driven (or other protein based algorithm) alignment to a reference species in order to identify ORFs and coding SNP markers. The shorter Illumina reads will produce a more fragmented transcriptome but the higher coverage will allow for the identification of more markers. The data presented here had been sufficient to design an initial panel of 96 markers. In addition to the paternity and linkage map projects, an ultimate aim of this study is to address the more general question of how heritable traits are in a wild population. We can estimate the additive genetic variation in wild animals via laboratory quantitative genetic studies, but these may fail to reveal what heritabilities will be in natural situations.

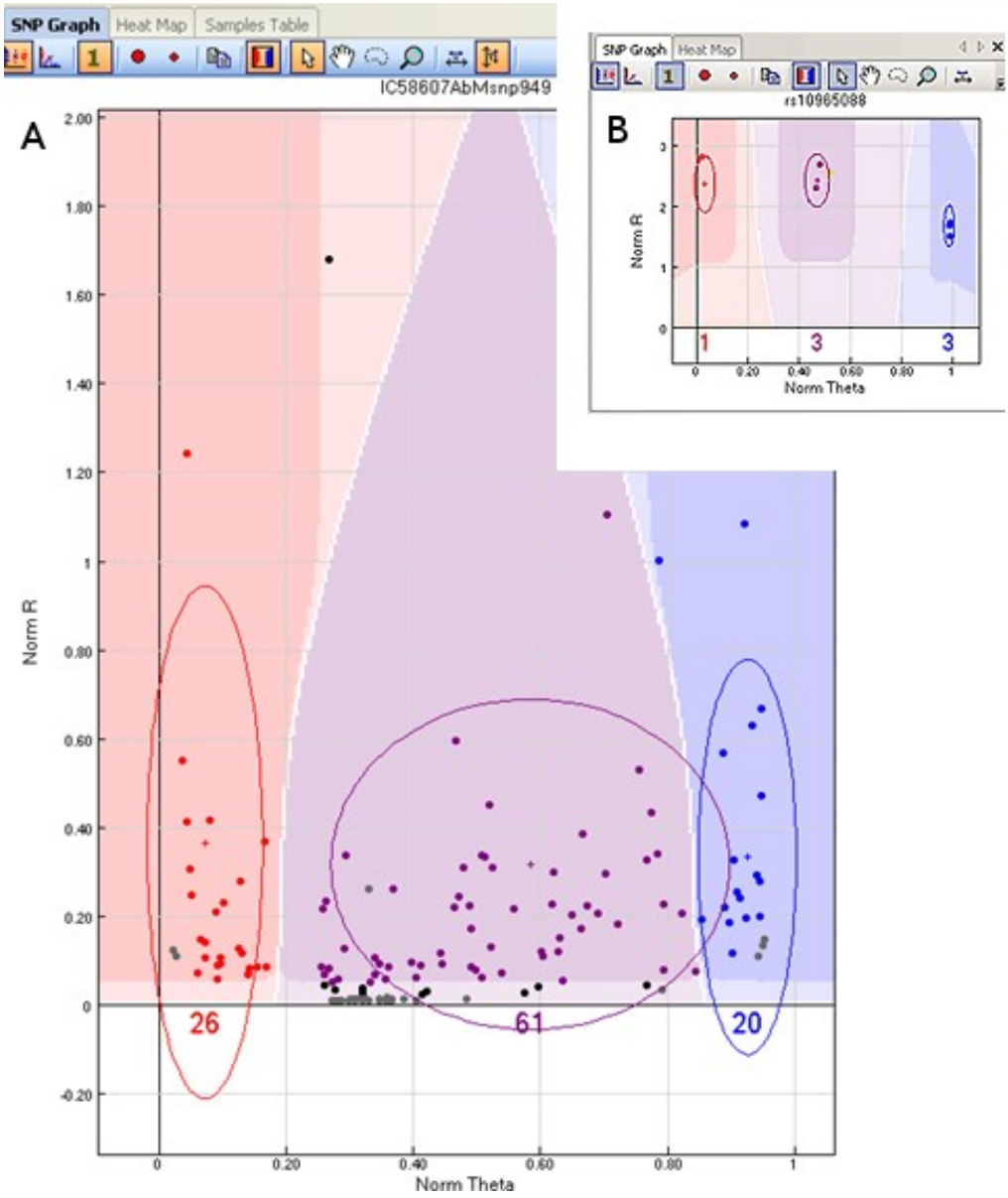


Figure 5. A) Illumina GenomeStudio screenshot of a single *G. campestris* locus (IC58607AbMsnp949) with fluorescense intensity correlation (norm R) on Y-axis and allele frequency on X-axis. Each dot represents a sample including positive controls, individuals from a pedigree and 32 individuals from a natural population. The software defines clusters for each genotype: red and blue are homozygous for each of the two alleles, purple is heterozygous. With high quality DNA the clusters are compact and not overlap. Black dots are animals whose genotype was not called. The grey dots, low on the Y-axis, are the natural population samples. Their low normalized R values are indicative of poor PCR amplication, probably due to low quantity and/or quality of DNA template. B) an example of how a typical GenomeStudio analysis looks like (provided by Illumina).

Manduca sexta and nicotine detoxification

Introduction

Manduca sexta is an important pest species on tobacco and partly due to its ease of collection, rearing and size is used as a model system for a number of research fields in functional biology including xenobiotic detoxification. Currently, the bulk of research on this field has been centered around cytochrome P450 (P450) enzymes and their activity in relation to detoxification of nicotine (a main defense compound of the insect's Solanaceous hosts) (Snyder et al. 1995; Stevens et al. 2000). Previous work has shown that P450s are abundantly expressed at the frontier of a larval's encounter with xenobiotics: the midgut (Feyereisen 2006). There is, therefore, good reason to consider P450s as good gene candidates for understanding how nicotine is metabolized by M. sexta and understanding how Manduca has adapted to be able to detoxify such a toxic compound. If, however, we wish to study the evolution of such a trait, we would have to be able to reconstruct the pathway in its entirety and P450s may have central but not solitary role. It is of interest, therefore, to use an exploratory Large Scale experimental approach in order to acquire more data on what is happening to an insect when challenged with nicotine and survives. This approach is a first step in a larger systems oriented approach and it can be addressed from a multitude of levels.

For a non-model insect it is perhaps easier to begin gather transcriptome based data. A reference and subsequently deep-annotation and data dissemination is a vital first step. As in genome projects, it is important to note that transcriptome assemblies are dynamic entities and the dissemination system used should, therefore, be able to handle the possibility of re-clustering as novel data become available. Upon having such a reference transcriptome we can proceed to use it in the place of a reference genome to conduct transcriptome-based surveys. One such potentially powerful experiment attempted here is to examine the levels of expression across the entire transcriptome and investigate how it alters with one or more experimental variables. The NGS technologies have made such examination statistically meaningful and low-cost. One method is RNA-seq, which aims to examine the entire mRNA transcript and investigate how its structure alters between experimental treatments. In the conxtext of expression levels, another method, deep-SAGE (SAGE from the older Serial Analysis of Gene Expression technique and deep due to sequence coverage) is forgoing coverage of the message in favour of sequencing depth. In this technique, the mRNA pool is bounded to avidin beads via a biotinylated poly-A tail, it undergoes a restriction digest and ligated adaptors limit sequencing at the tag adjacent to the 3'-most restriction site. This effective mRNA fractionation vastly increases the sequencing depth allowing for a more sensitive statistical

treatment. There are caveats however: first an mRNA without the restriction site will never be surveyed and second, an allele harbouring a mutation at the restriction site would give false numbers. The first issue can only be overcome with the use of multiple restriction enzymes, where as the importance of the second issue can be diminished by the use of inbred colonies and/or the increased number of biological replicates. In this case study, a deep-SAGE experiment was therefore performed with three biological replicates on the mRNA population of a single tissue in order to compare how the provision of nicotine alters gene expression at the insect midgut.

Methods

deep-SAGE data generation and preprocessing

Transcriptome sequence data was downloaded from the Short Read Archive (SRA) of NCBI and were originally derived from larval midgut tissue of multiple individuals (Pauchet, Wilkinson, van Munster, et al. 2009). The raw data sequences were processed with est2assembly v.0.99 and a reference transcriptome generated. In order to investigate gene expression differences between naive M. sexta individuals and those feeding on nicotine, 5th instar individuals reared on artificial diet were randomly assigned to nicotine-feeders or non-nicotine feeders and allowed to feed for 24h. Three individuals from each treatment were selected for the downstream application. From each biological replicate, larval midgut cDNA was generated as for the reference transcriptome. At the GenePool facility of the University of Edinburgh (http://genepool.ed.ac.uk), tagged ca. 17 bp cDNA restriction fragments were generated with NiaIII and MmeI following the standard Illumina protocol “Preparing Samples for Digital Gene Expression-Tag Profiling with NlaIII”. Briefly, it entails re-synthesising the anti-sense cDNA strand so that it is bound to oligo(dT)-beads; restriction with NiaIII and purification, ligation of adaptors introducing a sequencing primer and MmeI restriction site, restriction with MmeI to generate 17 bp tags; ligation of sequencing adaptors (i.e. complementary to the the oligos found in the flow-cell); PCR-driven enrichment and finally sequencing in a Illumina Genome Analyzer II using a 35 bp run protocol.

For each sample, a single lane of 35 bp Solexa was run to generate approximately 1,000,000 sequences for each sample. Custom scripts (deep_sage and digitra) were created to aid with preprocessing and analysis: each sample dataset was 3' trimmed to 18 bp to remove the adaptor sequence; the CATG restriction site was added to the 5'. Quality control was accomplished via the FASTX toolkit (available from http://hannonlab.cshl.edu/fastx_toolkit). At this stage, I removed large homo-oligomers which would confuse the aligner and unique sequences in the run were

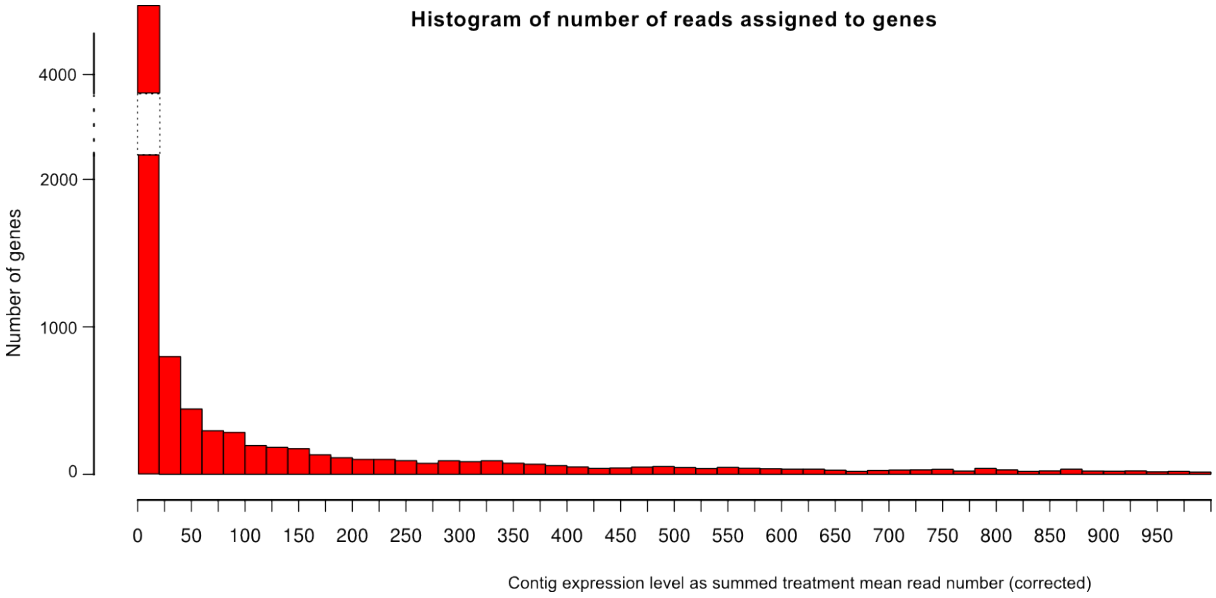


***Figure 6**: Frequency histogram of deep-SAGE Illumina reads aligning to M. sexta reference contigs after correcting for library size. Based on this graph, a cut-off of 100 reads was used to select the reliable dataset for downstream analysis.*

counted to give an overview of the number of unique sites surveyed. Subsequently, each sample was aligned to a non-polymorphic reference transcriptome using the BowTie aligner (Langmead, Trapnell, et al. 2009) allowing for one mismatch in a 10 bp 3' seed. The aligner was parameterized so that alignments with an average quality reduction below 50 (on the Phred scale) were rejected. Reads which did not map where further trimmed by 1 bp from the 3' end and re-aligned. In both alignment attempts, only one alignment per read was allowed, with the best one kept using the -stratum and -best options of BowTie. The digitra script outputs a tab-delimited file across all samples which includes statistical operations performed in R. Annotations of contigs was acquired from the InsectaCentral database.

Statistical analysis of deep-SAGE data

First, counts of aligned Solexa sequences were measured for each reference contig. Normalization for gene length was not performed: unlike an RNA-Seq protocol, deep-SAGE is dependent only on number of restriction sites, which we assume are unchanged between individual insects (i.e. no point mutations at CATG sites). I did investigate normalization with two methods. First, a standard method using the mean number of reads normalizes the number of reads i in lane l by multiplying I with k, a lane-specific ratio which is the mean number of sequences across all lanes / total number of sequences within lane l. In this dataset, kappa was recalculated after C3 was excluded (see

results). The second normalization approach was via the TMM method as proposed by (M. D Robinson and Oshlack 2010). I also estimated the fold-difference between sample groups, taking the group mean. Contigs which had fewer than 10 reads (of ca 10,000,000) in all samples were excluded from the analysis as i) they could be misalignments or collapsed repeats ii) statistical testing of difference between treatments would not be appropriate due to the high number of multiple testing instances. In both cases, in order to assess statistical significance, the use of t-test was explored initially but not used because of the low number of replicates and violation of the assumptions (non t- or even normal-distribution). I tested, therefore, whether the treatment factors affected the data's fit to a general linear model following a quasi-poisson distribution. The quasi-poisson distribution has been shown to be applicable in highly dispersed data such as those in this experiment (Marioni et al. 2008; M. D Robinson and Oshlack 2010) and was the one giving the best fit to the data. An ANOVA F statistic was used to test for significant difference in fit of the null vs an alternative hypothesis that the samples where from different, non-overlapping distributions. A second approach was also tested as proposed by (M. D Robinson and Oshlack 2010; M. D Robinson and Smyth 2007): a Fisher's exact test between the normalized means of the two samples.

The resulting data were sorted by p-value and fold-difference in order to allow experimenters to first test (in the wet-lab) the genes which are most significant at a specific False Discovery Rate (FDR). The FDR approach (as implemented by (Benjamini and Hochberg 1995)) was estimated in R using the p.adjust function; it was also computed via two Bayesian approaches as implemented in the fdrtool package of R (Strimmer 2008a) and the qvalue package as implemented by (Storey and Tibshirani 2003).

Results and discussion

Statistical treatment

From aligning to the reference transcriptome, 11,005 contigs were represented in the Illumina dataset. This dataset had biological replicates which allows for a more thorough statistical treatment in order to find which of these gene objects had evidence for being differentially expressed by the induction of nicotine. The dataset showed high dispersal and therefore a non-poisson distribution had to be used: either a negative binomial or a Poisson with the gamma parameter (quasi-Poisson). I compared the two methods of estimating p-values: the GLM approach on normalized counts using a quasi-Poisson distribution and the Fisher's exact test of the normalized means assuming a negative binomial distribution. Despite some differences, the methods correlated but the latter method was

chosen for the presentation of the results. The GLM approach is more conservative but can useful in the future when there are more than two treatment groups and the Fisher's exact test would be limited to time-consuming pairwise comparisons.

Further investigation showed that the number of tests can be substantially reduced, increased thus the power of the approach. In total, 253 contigs had an irregular expression level (one of the replicates only showed expression) and they were not tested. Before testing for significance, I excluded contigs which had an exceptionally low number of reads; a histogram of the summed, mean expression level across treatments shows that a large number (6,157) of contigs have less than 100 reads (Figure 6). Therefore, a contig had to have at least 100 reads in every sample or at least 500 reads across all samples. This was done to avoid a) false alignments due to mismatches b) allow for more robust statistics at the expense of losing some possibly differentially expressed genes which have a basal level of expression. As the biological phenomenon is associated with detoxification, we do not expect that expression levels will be at the basal level.

Global visualization

A global visualization approach via a heatmap was used to detect that one sample was an outlier and should be discarded prior the statistical analysis (Figure 7A). A non-clustering heatmap (Figure 7B) was generated to get an overview of expression and visually verify that the expected number of differentially expressed genes (Rajaram and Oono 2010). A non-clustering approach was used because like in an EST analysis, clustering tends to remove information from the data, especially with a small number of replicates: the clustering of two data points influences the clustering of all the subsequent ones. As an effect, an incorrect clustering event (a bifurcation in the dendrogram) will influence all subsequent ones. In phylogenetics, bootstrap analysis allows us to evaluate the robustness of each clustering event but this is not, however, done in most hierarchical clustering implementations in genomics (it would computationally challenging with thousands of features). Instead, I used multi-dimensional scaling (MDS) to attribute weights to each feature. MDS is a dimension reduction algorithm similar to PCA which has been shown to perform better in such scenarios as microarray analysis (Thalamuthu et al. 2006). From Figure 7B, it can be seen that there are two sets of genes are consistently overexpressed in 2 of the 3 treatments and not in controls. A third set was underexpressed in relation to the controls.

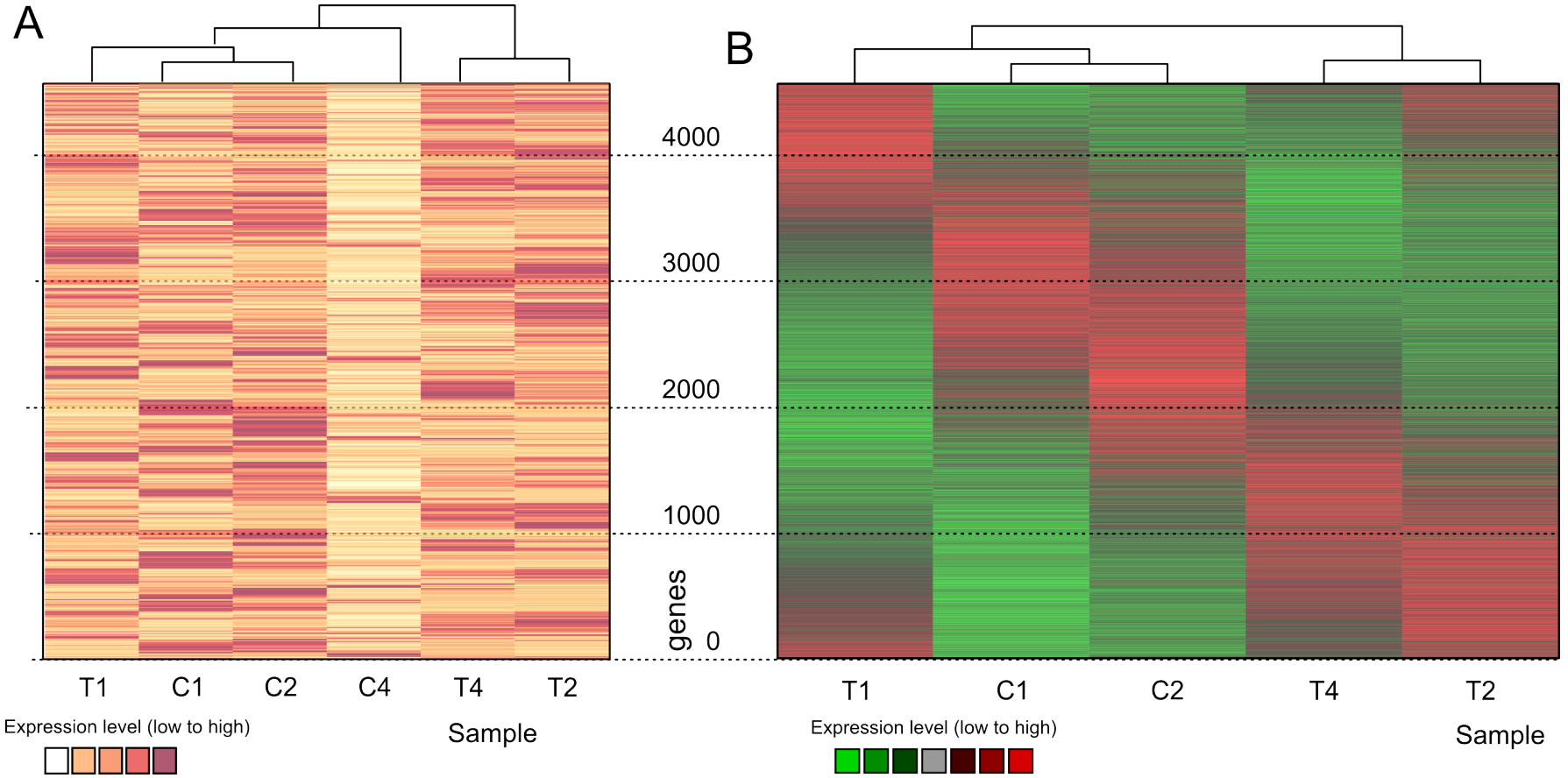


***Figure 7**: A) Heatmap of all M. sexta deep-SAGE samples after hierarchical clustering. Sample C4 is an outlier due to sample or sequence quality and was not included in downstream applications. B) Heatmap of remaining samples, sorted using MDS weights. Shows the expected number of genes differentially expressed and hints the distribution. Both more controls and treatment samples seem to be needed.*

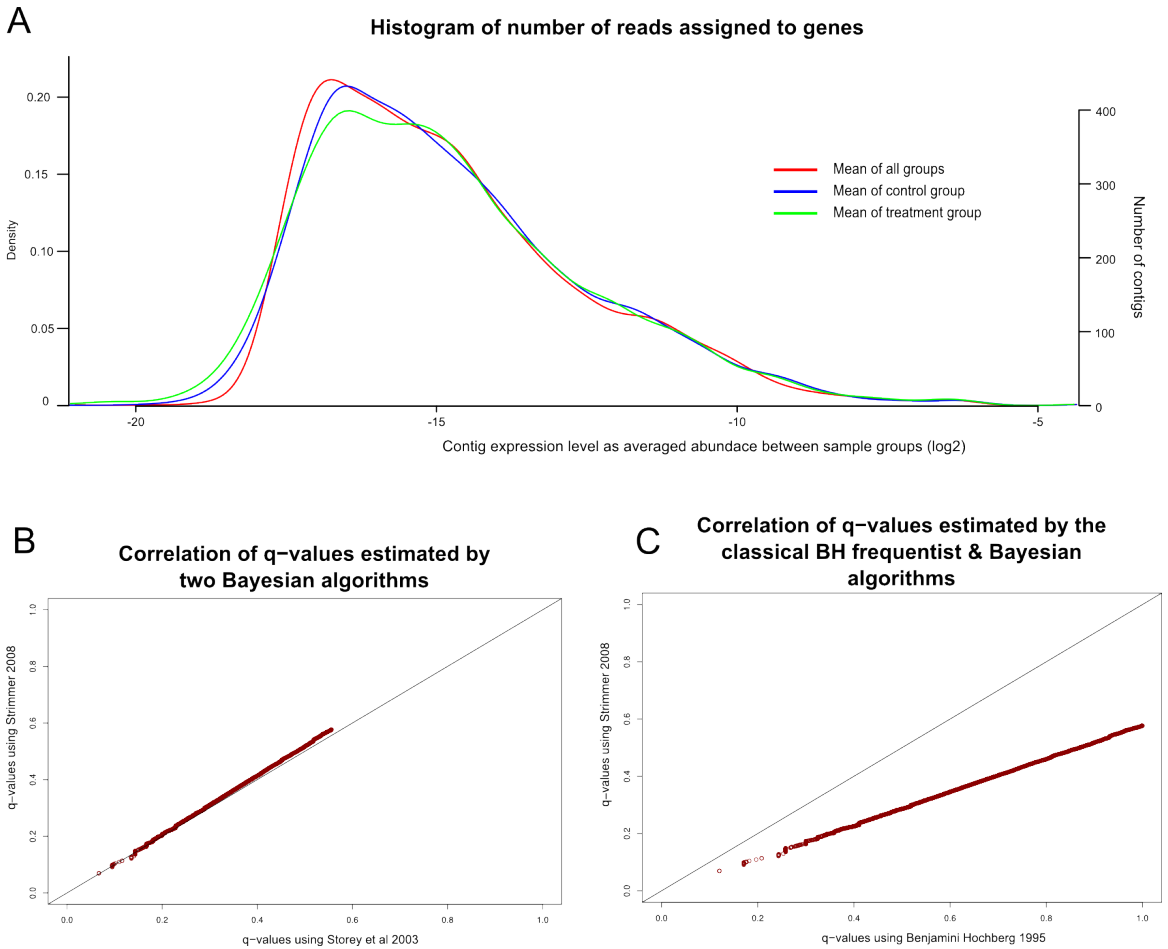


***Figure 8**: A) Distribution of contigs in relation to the expression level (i.e. abundance) when average within and between sample group; compare with Figure 10. The three abundance distributions are similar and therefore abundance does not influence if a gene is estimated as significant. B/C) Q-Q plots showing correlation of FDR q-values obtained by three different methods: B) Storey et al 2003 vs Strimmer 2008 and C) Benjamini – Hochberg (BH) 1998 and Strimmer 2008.*

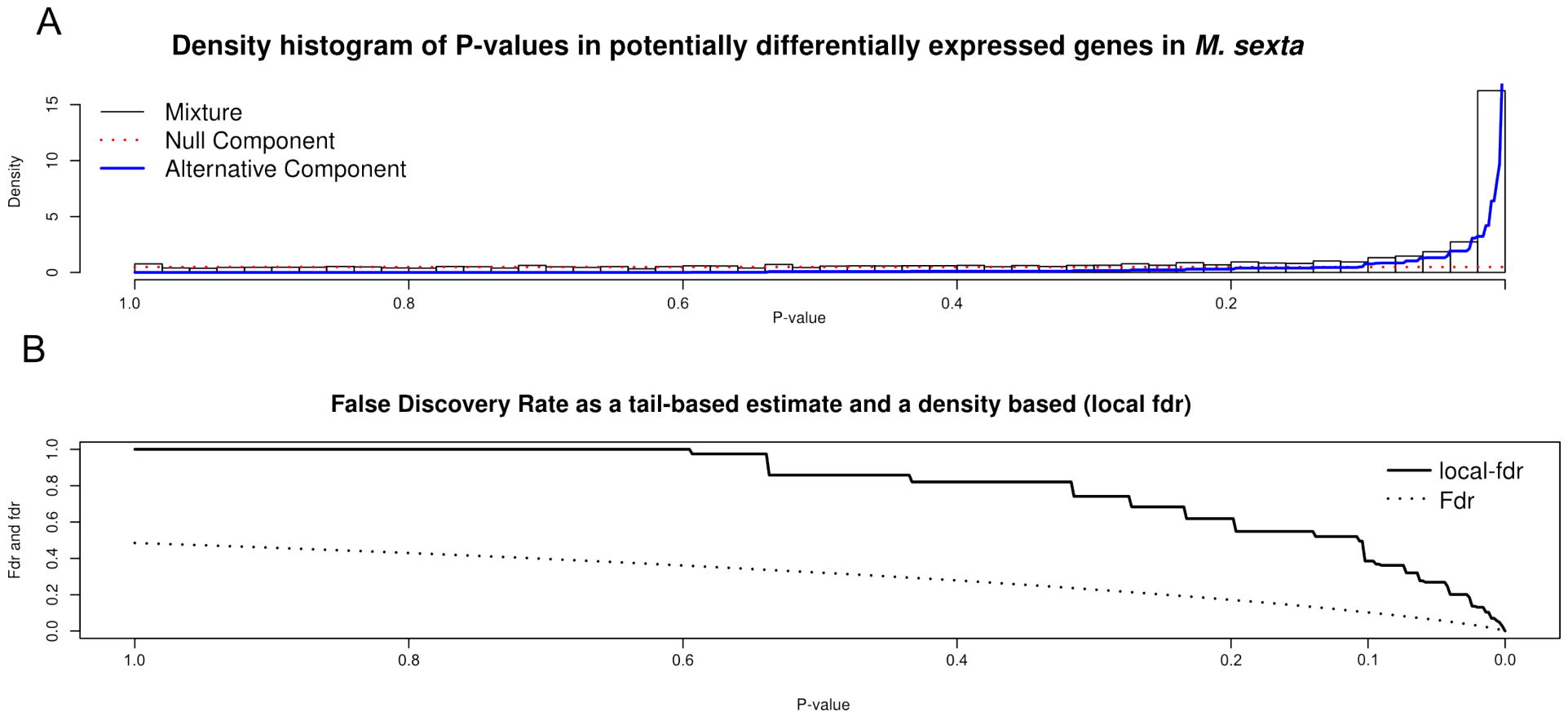


Figure 9: *A) Histogram of p-values and fit of distribution expected under the null and alternative models for M. sexta deep-SAGE contigs. B) Q-values estimated using standard FDR and local FDR.*

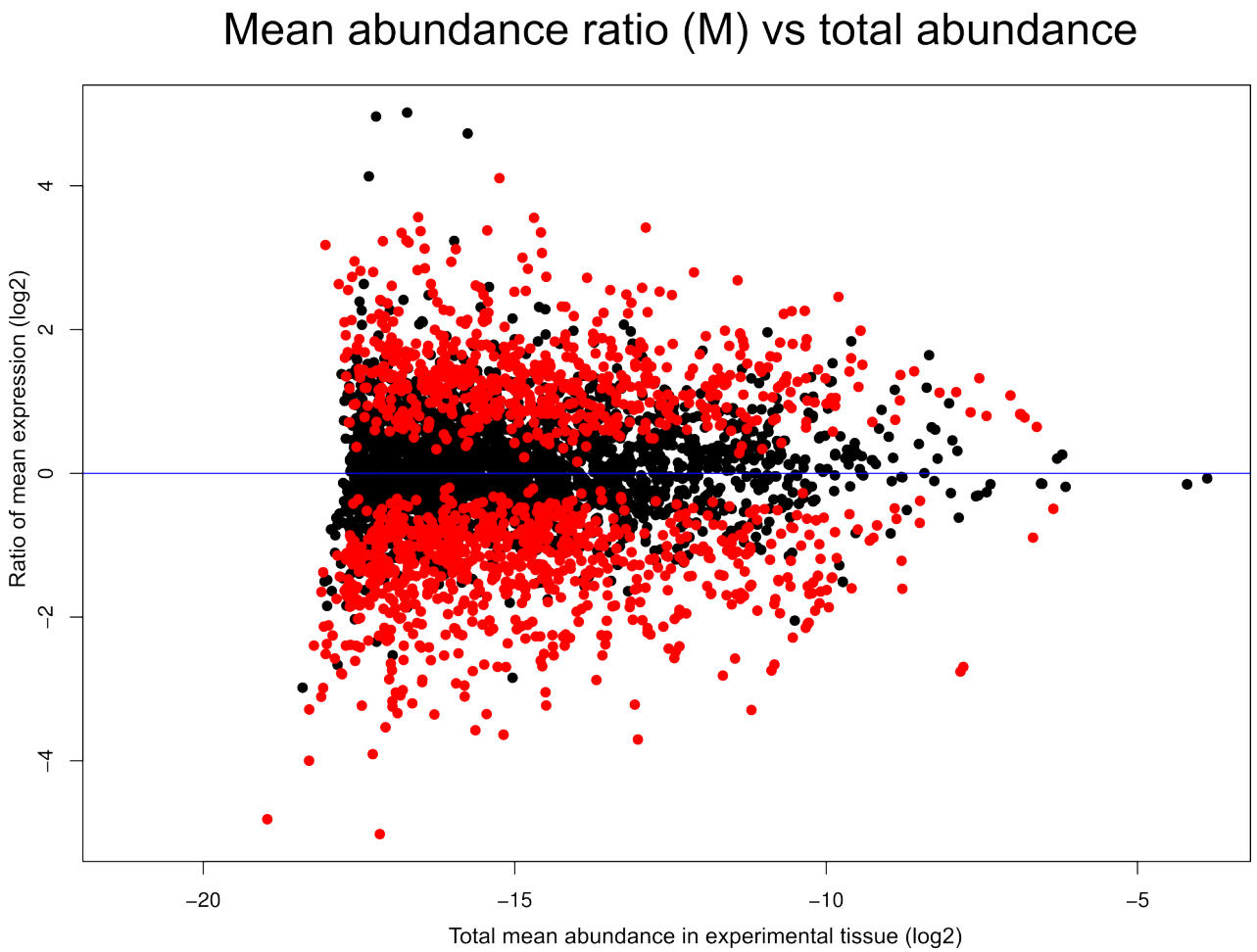


***Figure 10**: Plot of M. sexta deep-SAGE expression ratio of means from control vs treatment groups (log2) versus total abundance as sum of the two group means (log2). Black spots shows comparisons judged non-significant and red spots are significant at an FDR of 1%. Compare with Figure 7*

Differential expression across treatments

The estimation of significance approach showed that 1,549 contigs were significant at an error level of 5% (p-value <0.05) with a minimum p-value of 0. The distribution in relation to fold change is shown in Figure 8A. This approach does not take into account an important fact: 3,901 tests were run and each test has a 5 % probability of showing as significant due to chance. The common approach in such cases is to use Bonferroni's correction. This, however, is over-conservative in experiments with such high number of tests: only p-values less than 1.3e−5 would be significant. Further, not all tests have an equal chance of being incorrect: the smallest the p-value, the less likely the result is due to chance. Worse still, many bioinformaticists use an ad-hoc cut-off such as 1e-6 (personal observation). A more appropriate approach is to implement a false discovery rate (FDR) approach and therefore relate each p-value with an FDR q-value: the proportion of contigs (and not as the probability as some authors write) with this or lower p-value being false positives.

Estimating q-values using the classic BH FDR algorithm is conservative and improvements using a Bayesian approach have since been implemented (Strimmer 2008b; Storey and Tibshirani 2003). The calculation of the prior, the expected number of tests which are non-significant, was estimated using simulations via two methods as proposed by Strimmer 2008 (Strimmer 2008b) and Storey and Tibshirani 2003 (Storey and Tibshirani 2003). The results was 0.485 and 0.481 respectively. A scatter plot of the q-value sets estimated with these two methods showed that they correlated well (Figure 8B). This was not the case when compared with the far more conservative approach implemented by the frequentist statistics algorithm of Benjamini and Hochberg 1995 (Benjamini and Hochberg 1995) (Figure 8C). Indeed, the BH gave few significant contigs and this is because it sets the proportion of non-significant contigs to equal to 1, i.e. it assumes that all tests would be non-significant. Such an assumption is clearly wrong in the particular biological scenario: upon treatment with a plant defense compound we expect a number (and probably a large one) of genes to be differentially expressed. Indeed, the approach used here can provide us with an estimate of the expected number of differential genes (Figure 9A).

The FDR approach is a powerful solution which solves the multiple-testing issue and approaches a problem intuitively: by estimating the q-values we can select our level of allowed false positives (hence the name False Discovery Rate) depending on the application. To illustrate the approach, let us assume that we are comfortable with up to 5 genes being false. In that case, the first 847 genes would be the suitable candidate set, after ordering the genes with increasing p-values (360 if only allowing for upregulation). Likewise, we can ask how many genes are there if we can tolerate a

specific false discovery rate. Allowing, therefore, for a FDR of 1 % (1 % of contigs called significant will be false positives) we find that 941 genes are differentially expressed between the treatment and control and therefore 9 to 10 genes will be false. Of 941, 407 were upregulated and 534 were downregulated in the treatment. The fact these two number are similar has been affected by the TMM approach used which is designed specifically to address the phenomenon of expression level ratios deviating from zero.

In addition to the FDR approach used above, I also calculated the local-FDR (or abbreviated as fdr; see Figure 9B). A FDR q-value is the proportion of contigs which are false positives if we accept all contigs with a given or lower p-value, the local-FDR more appropriately informs us about the probability this specific contig is false (Storey and Tibshirani 2003; Efron and Tibshirani 2002).Thus, for example, the 941st most significant contig (the last one in the chosen cutoff) IC7130AgEcon1869 (protein IC7130AgApep5481) has a p-value of 0.005, a q-value of 0.01 but a localFDR q-value of 0.056 (i.e. higher than 0.05). This translates to: i) in the first 941 contigs, 9.4 genes (0.01 x 941) are false positives; ii) IC7130AgEcon1869 has a chance of 5.6 % of being a false positive.

Annotation

The number of genes that have significant differential expression is, however, too high to allow for a wet-lab biologist to test each gene explicitly. We have to, therefore, annotate the genes in such a way so that a wet-lab biologist can use biological knowledge to derive candidates. Further, this approach can be used to give a global birds-eye view. Our next step is, therefore, to ask if these genes are related in some biological dimension. Via est2assembly, each contig was related to a protein sequence which was stored in InsectaCentral along with annotations such as GO and KEGG terms. We can thus compare the distribution of these GO and KEGG terms as found in the entire transcriptome (“reference space”) to those found in the significant “subset”. For example, we could expect a variety of P450 enzymes to be differentially expressed (Snyder et al 1995 (Snyder, Walding, and Feyereisen 1995) but see Stevens et al 2000 (Stevens et al. 2000)) but in reality only a small proportion of identified P450s are present in the subset (22 in the reference space versus 4 in the subset). This fits with the data presented by Stevens et al (Stevens et al. 2000) which shows that there is an array of P450s whose expression profiles can be surprisingly specific to certain xenobiotics. Indeed, we identify CYP4M1 in the subset (InsectaCentral/GenBank identifier: IC7130AgEcon1415/L38670). It is upregulated in the treatment by 1.84 fold and is ranked 79th in the FDR q-value rank. The other gene identified by Snyder et al (Snyder et al. 1995) is CYP4M3

(IC7130AgEcon1615/L38672) and has no evidence for downregulation (p-value is 4e-01). For this gene there was significant variance within sample types: the two controls had 292.91 and 851.26 counts and the three treatment had 644.41, 259.07 and 242.9. Indeed, if only the first treatment/control set had been used these gene would have been shown as differentially regulated. Additional samples can assist in determining the correct status of such cases.

The number of contigs significantly overexpressed in individuals challenged with nicotine is roughly 8 percent of the total number of contigs that had some representation in the Illumina dataset and about 24 percent from the trimmed subset (i.e. after removing contigs with a low number of counts) but a number of caveats exist. Because of the large number of contigs discarded during trimming, it is possible that insufficient coverage has been achieved and that this experiment might benefit from an additional sequencing run. Alternatively, mutations that exist between experimental individuals can result in i) spurious alignments and/or ii) a gene not representated accurately in all samples. This is a caveat with all deep-SAGE experiments on polymorphic, non-inbred organisms. Genes which had no Illumina reads aligning to them do not necessarily indicate a lack of expression: absence of a CATG restriction site would not allow for this gene to be sampled. One has to keep in mind, however, the statistical limitations of this method when a small number of individuals is used to test for significance. An estimation of the distribution is undertaken with a few data points which is problematic for genes with generally low expression. The actual shape of the distribution is unknown in both the actual sample and globally in the organism - i.e. we are not aware how each gene's expression fluctuates spatially and temporally - but can be inferred only from the data. The estimation of significance is made thus more robust with an increased number of replicates. The Illumina technology has been claimed not to require technical replicates (Marioni et al. 2008) even though it needs normalization. As mentioned, biological replicates, on the other hand, introduce noise, especially if they are not derived from the same genetic background, have high heterozygosity or contain contamination from other tissues. Indeed, a future design approach would be to use the offspring from a single-pair mating - with the caveat, however, of potential family-specific effects which can be detected with the use of multiple families.

Even though the above issue applies to experiments in any organism, additional considerations apply to non-model species. First, the reference transcriptome was generated from a single tissue of a narrowly defined developmental stage. Any candidate genes not captured in that sample used for generating the reference transcriptome will not be detected during alignment of the deep-SAGE approach. Further, it was automatically determined using est2assembly and the MIRA2 assembler; i.e. it was not manually curated. This results in certain contigs showing redundancy, often even

manual curation cannot collapse them without a full length cDNA sequence derived experimentally. Even though the reference space might benefit from curation, this should not alter the counts: the approach used ensures that each read will align to the best contig encountered. Barring therefore mutations in the 18 bp sequence downstream of the CATG site, then contigs with lower numbers will be favoured when a gene has a redundant representation. On the other hand, both the assembler and the aligner software were built for model species, namely bacteria or low heterozygosity human or fruitfly samples. The MIRA assembler has non-model species improvements if a correct set of parameters are chosen (see est2assembly chapter) but heterozygosity issues still exist due to the high number of individuals (i.e. alleles/chromosomes) in the sequencing pool of the reference transcriptome. The aligner does not allow for IUPAC codes to denote polymorphisms; computational efficiency is achieved by an implementation of the Burrows-Wheeler transform which requires 4 character states. One has to, therefore, convert any polymorphic sequence to a fixed allele. Such a procedure should not, however, bias the counts unless all individuals in a treatment had a different allele for one or more SNPs near the CATG sites. The issue of mutation of a CATG site itself is more problematic. If some individuals were heterozygous for a particular CATG site, the number of reads sequenced for this allele would be halved. One potential solution is to repeat the experiment with an enzyme using a different restriction site.

Addressing the above concerns could be helped via an experimental validation of the best candidates and a small panel of randomly picked negatives. Full-length cDNA sequencing and a qPCR methodology would be the most thorough but time-consuming approach. The former, coupled with manual curation, will allow us to re-align and re-estimate candidate genes, in light of a partially curated transcriptome. Indeed, if the experimental animals are used in the qPCR, then the issue CATG mutation could also be investigated.

Conclusion

Overall, it is obvious from the MA plots (Figure 10) that this experimental setup is powerful enough to detect very small differences in expression levels. Indeed, the deep-SAGE approach is at least as powerful as a similarly designed microarray study and it lacks the uncertainty bias associated with the image analysis step. It is also more powerful than traditional SAGE experiments due to larger amount of data. Regardless of species or system used, all differential expression studies depend on capturing the differential expression event in the experimental animals at the time of tissue harvesting. Due to cost, the particular experiment does not include a time-series design and therefore it is assumed that all experimental animals were at a functionally equivalent

developmental stage. For the particular experiment in Manduca sexta, this would result in an increase of false-negatives due to the increase in within sample-group variance. The high number of candidates already found, shows that candidates are expressed in this developmental window (or possibly constituentively) and therefore the results are sufficiently robust. Should we, however, wish to construct a complete developmental network, we would perform a thorough sampling in both a spatial and temporal manner. In a first instance, this approach would utilize the Gene Ontology enrichment approach to identify metabolic processes or functions. In a second instance, it would provide annotations for genes with annotation function. Indeed, a detailed deep-SAGE analysis of transcription for a number of treatments of a particular theme would be essentially a cost-effective alternative to painstaking biochemical functional analysis. In concert with a thorough manual curation of the transcriptomes of this (and perhaps closely related species), a true picture of the membership and evolutionary dynamics of pathway elements would, for the first time, begin to emerge and current theoretical models tested and improved.

Papilio species and genetics of wing pattern variation

Introduction

The two swallowtail butterflies, Papilio glaucus (eastern tiger swallowtail) and Papilio dardanus (African swallowtail) occur in N. America and Africa respectively and represent some of the most famous examples of Batesian mimicry in butterflies. Batesian mimicry occurs when an aposematic toxic model species has its aposematic signal mimicked by another species which does not produce an avoidance-reaction upon eating by a predator. In order for the aposematic signal to be maintained, the non-toxic mimic is established in a lower frequency than the toxic model. This contrasts with Müllerian mimicry where both the model and the mimic are both significantly toxic to the same predators and can, therefore, reinforce the aposematic signal. These Papilio species exhibit aposematic wing colour-patterns but the signal is limited to the female sex: the males have a distinct non-mimetic colour pattern. In the case of P. dardanus females, there is a great diversity in the types of colour patterns, because several different species serve as models. Since the males are monomorphic, colour pattern morphs but not races have evolved and are maintained throughout sub-saharan Africa (Clarke and Sheppard 1963). For P. glaucus females only one mimetic morph has been recognized, which mimics the pipevine swallowtail Battus philenor.

Genetic work by Clarke and Sheppard has established that the major locus controlling the P. dardanus colour pattern is autosomal and at least 10 alleles are known (R. Clark et al. 2008; Clarke and Sheppard 1963). Similar work by Scriber shows that the mimetic locus of P. glaucus is located

on the W (Y) chromosome with possible epigenetic complications or epistatic effects linked to the Z chromosome (Scriber, M. H. Evans, and Ritland 1986; Scriber, R. H. Hagen, and Lederhouse 1996). Current work on both species is focused in determining the loci controlling the colour pattern. For P. dardanus a classical linkage mapping approach is being utilized (R. Clark et al. 2008) but for P. glaucus this is not possible due to the fact that the W-chromosome is female specific and undergoes no recombination. A cytogenetic method has, therefore, been undertaken (Fukova, U. Exeter) in conjuction with work presented herein to provide candidate loci involved in wing pattern formation, melanism or sex-biased differences using a transcriptomic approach. This approach has been complemented by similar work from another Papilio species, P. xuthus, where larval markings including melanism were studied (Shirataki, Futahashi, and Fujiwara 2010; Futahashi, Banno, and Fujiwara 2010; Futahashi and Fujiwara 2006; Futahashi and Fujiwara 2005; Futahashi and Fujiwara 2007). It's not known, however, which genes control or activate this pathway and also if these are similar across the genus.

Methods

Experimental animals

P. glaucus female offspring of field collected females (Pennsylvania) by Aardema and Scriber (Michigan State University). The butterflies were mated with male offspring originating from different families. Offspring coming from these crosses were used for preparation of wing disk cDNA libraries. Larvae were fed on Prunus serotina (black cherry) leaves in laboratory temperature at Princeton, USA, under natural light (14-15 L : 9-10 D) and without humidity control. P. dardanus polytrophus morph hippocoonides larvae originated from wild caught mothers from Arabuko Sokoke forest (Kenya). Offspring of these females were shipped together with their host plant to the United Kingdom. Larvae were kept in laboratory conditions, under natural light and without humidity control.

cDNA libraries

The P. dardanus RNA samples were generated from the pre-pupal wing discs of 7 males and 7 mimetic females (P. dardanus) by Dr Fukova (U. of Exeter). For P. glaucus, pre-pupal wing discs from one male and mimetic female were used. For each sample, cDNA was generated using the SMART IV kit (Invitrogen) after the RNA of each species was pooled. Prior to sequencing, each cDNA pool was normalized with the Trimmer Normalization protocol (Evrogen). For P. dardanus,

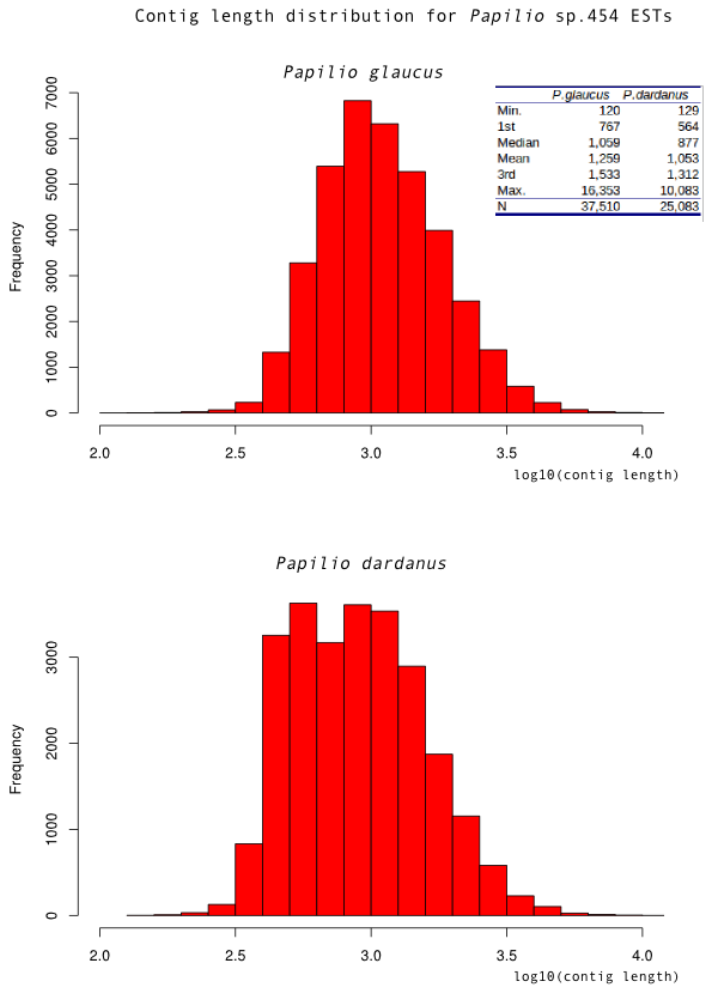


***Figure 11**: Distribution of contig lengths (log10) for P. glaucus (A) and P. dardanus (B).*

Next-Gen 454 GS-FLX Titanium sequencing was performed at the Advanced Genomics facility at the University of Liverpool (http://www.liv.ac.uk/agf) and for the P. glaucus dataset 454 GS-FLX Titanium sequences were provided from the Human Genome Sequencing Center at the Baylor College of Medicine (c/o Dr. Rui Chen). For the deep-SAGE experiment in P. dardanus, the same RNA extraction was used to produce two pooled samples: a mimetic female and a male. Note that the genotype of the males is unknown but the phenotype is always non-mimetic. No technical or biological replicates were performed in this case.

Generating a reference transcriptome & melanic pathway annotation

The EST reads or contigs from P. glaucus and P. dardanus datasets were saved as a local database in Geneious (versions 4.8 and 5.1, Biomatters, Auckland, New Zealand) and used for BLAST searches with known members of the melanin biosynthesis pathway (using as references species D.

melanogaster and P. xuthus). The reference protein and mRNA sequence were downloaded from NCBI and imported into Geneious. The sequences were used as a queries to search the local database using BLAST. The top ten hits (e-value < 1e-30) were downloaded and used to build a global alignment with the nucleotide reference sequence. Frameshifts were corrected and a consensus sequence was built from the aligned contigs or reads of the target species (75% identity). Global alignments of proteins were constructed in MUSCLE. For identification, the Geneious tree builder was used for generating neighbor-joining trees under Jukes-Cantor distance model. Bootstrap was performed with 500 replicates. Multiple reference species were used for proteins showing only partial conservation of sequence with a contig from the target species (e.g. Ebony). In these cases, a Hidden Markov Model was built using HMMER (Eddy 2000). The predicted peptides of the target species were searched with hmmsearch using the 'gathering threshold cutoffs' parameter of HMMER. Curation of the ORF proceeded as above but it should be noted that one should be less confident of the quality of the sequence until it is verified by full-length sequencing.

Curation of reference BAC sequence

Assembled Papilio dardanus BAC contigs, obtained from GenBank and Dr H. Vogel, were pairwise aligned using dottup and dotmatcher from the EMBOSS package and a single contiguous sequence was reconstructed manually. The final, reference, BAC sequence was first annotated with: i) the transcriptomic data using est2genome, a gapped alignment program with intron/exon boundary recognition; ii) CDS models driven by KAIKOGAAS (Shimomura et al. 2004) as provided by Dr. S. Baxter (U. Cambridge); iii) sequence similarity matches (via BLASTx) with the Uniref50 database; iv) de-novo gene models produced by SNAP; v) repeat regions as identified by RepeatMasker; vi) regions shown to be differentially expressed in males and mimetic females by the Solexa digital gene-expression profiling. Illumina-derived tags were aligned to the BAC using the Geneious assembler with a word length of 5, maximum 1 b.p. gap and allowing up to 10 % of mismatches and 10 % of gaps. For the sample-specific tags, the number of allowed mismatches was increased to 15 %. The Maker pipeline was used to derive consensus for the annotations from (i) to (iv). Subsequently, all annotations, including the SAGE results, were loaded into a Geneious database and analyzed: the BAC was manually curated to identify repetitive regions, correct gene models and determine which of them were supported by transcriptomic evidence. The final, annotated, contig Geneious file is available upon request.

Statistical analysis of deep-SAGE data

The approach undertaken for this dataset was similar to the M. sexta dataset. As only two samples and no replicates were available, I used a Fisher's Exact test to judge significance of expression differences but after normalization for average read number, two different approaches were undertaken: i) test for significance each tag independently of any alignment to a reference transcriptome but used an alignment to a repeat-masked reference transcriptome for assigning annotation; ii) test contigs as a whole and only consider, thus, tags which aligned to the reference transcriptome. In the first case, tags were collapsed with the fastx_collapser program. For alignments to a reference, the BowTie program (Langmead, Schatz, et al. 2009) was used in the same manner as for the M. sexta dataset.

Results and discussion

Transcriptome sequencing

Sequencing and reference sequence generation was significantly enhanced compared to other datasets in this thesis, due to the use of the new generation of 454 protocol (GS-FLX Titanium). The P. glaucus produced the expected 1.2 M reads but P. dardanus produced only 940,005 reads, excluding 390,468 reads which failed. The cDNA generation protocol was the same, by the same individual but it is likely that the Baylor sequencing center used an adapted plating protocol (Jenn Schaff, North Carolina State University, pers. communication). In total, we acquired 37,510 and 45,792 contigs for P. glaucus and P. dardanus respectively with the former dataset showing a more normal distribution for contig length (Figure 11). Except lower coverage, the large number of individuals used for the library generation (14) is one potential reason for the inflated contig number. This directed us in generating the P. glaucus library using only 2 individuals since sufficient starting material was acquirable even from the target tissue.

Exploration of a new differential expression method using non-model species

The P. dardanus deep-SAGE dataset did not, due to cost, possess biological replicates but each of the 2 samples was the product of the pooling of 7 pre-pupal larvae which differed in sex and colour pattern (yellow-males vs melanic-females). A total of 10,338,899 and 12,322,165 reads (mean 11,330,532) were generated for the male and female samples respectively using Illumina sequencing. After generating unique tags and counts, the data were normalized as per the M. sexta dataset then statistical significance of differential expression was tested by considering i) summed

counts on contigs; ii) each tag independently.

Contig approach

As the reference transcriptome was developed without the trim_assembly approach used elsewhere in this chapter, I used this script to reduce the redundancy of the dataset to 95 %, resulting in 35,302 contigs. This next step was aligning the Illumina sequences to the reference transcriptome and considering counts across the entire contig. Of the total reads, 79 % of them aligned: 20,794 contigs were identified using the male dataset and 21,128 contigs using the mimetic-female dataset producing a total of 22,976 contigs identified using all SAGE data. Using a histogram approach as in the other SAGE datasets, I removed contigs where the row total was less than 100 counts. 11,753 contigs remained. After normalization, testing for significance and implementing a 1 % FDR, 2,183 genes were differentially expressed. Of those, 1,203 were upregulated in the female sample and 979 were downregulated. Of those, 56 contigs were female sample specific and 26 were male specific, i.e. had no (or less than 10) tags in the other sample (Table 6). Plotting of mean abundance versus relative abundance (M/A plot) shows the dispersion of the data (Figure 12A). Comparison with the M/A plot of M. sexta shows that for this dataset, significance is relying solely upon the ratio of transcription being above a certain threshold (which is dependent on relative abundance). Having more samples would allow us to detect significant data-points (shown in red) which are not robust (as in the M. sexta dataset).

Tag-centric approach

Instead of aligning tags to contigs and estimating differential expression based on summed tag counts, one can test for differential expression between each tag separately, i.e. without pooling to contigs. Annotation can still be transferred from an alignment. This approach may show regions of a mRNA that is differentially expressed (e.g. alternative splicing) even though the deep-SAGE method is not as a powerful as whole message sequencing (RNA-seq). A total of 470,071 unique tags were identified from a potential of 68,719,476,736 combinations (4^18). Searching the transcriptome for the restriction site CATG, I expected no more than 99,485 unique tags. To consider only useful tags and filter sequencing errors and repeats, a histogram was used to guide discarding tags with less than 50 total (normalized) counts in both male or female samples (Figure 13A-B). As a result, 23,288 unique tags remained of which 14,843 tags had a p-value equal or less than 0.05. In this case, the Bayesian estimation of FDR was uninformative (p-value was equal or less than q-value). This is likely to be because the estimation of the prior is not robust when too

many tests were significant (the prior was estimated to be 0.26 ~ 0.28 i.e. most genes are expected to be differentially expressed). Thus the BH implementation of FDR, which sets the prior to 1, was used. At an FDR of 1 %, 6,471 genes were significantly differentially expressed. A histogram shows that many of these genes are expressed at very low levels (Figure 13C), a second cutoff at 250 counts was therefore used. After recalculation of the FDR, this resulted in 4,445 genes with a p-value equal or less than 0.05 and at an FDR of 1 %, 2.226 of them were differentially expressed. As expected via the normalization, half were upregulated in the mimetic female sample (1,114) and half were downregulated (1,112). Estimations using means normalizing by just mean library length (i.e. without trimming as conducted by the edgeR package) showed no differences. The M/A plot for tags (Figure 12B) shows the extreme dispersiveness of the data when tags are considered. It is likely that with tags we will need to consider a more conservative FDR.

Table 6 Results of aligning contigs & tags with reference transcriptome

	Tags		Contigs	
	Female-sample	Male-sample	Female-sample	Male-sample
Sequences in lane/contigs aligned	12,322,165	10,338,899	21,128	20,794
Upregulated within an FDR of 1 %	1,114	1,112	1,203	979
Sample specific	71	56	56	26
Sample specific tags aligning to reference contigs	34	39	N/A	N/A
Sample biased tags aligning to reference to contigs	751	661	N/A	N/A

Differential expression can be found between samples which both express a particular message, these candidates are sample-biased. Further, tags may be limited to one sample, i.e. be absent from the other. These are sample-limited tags. There were 43 and 41 tags specific to females and males respectively (Table 6). This was calculated using only tags which zero counts in the other sample. Using a histogram, I defined that up to 2 sequences in the other sample could be considered as misalignments, increasing the number of specific tags to 71 and 56 respectively. For each of the these tags I aligned them against the reference transcriptome in order to assign putative annotation. Of the sample limited tags, 34 of the female sample and 39 of the male sample matched a reference contig. Of the sample biased tags, 751 (69 %) and 661 (65 %) female and male samples respectively, aligned to the reference.

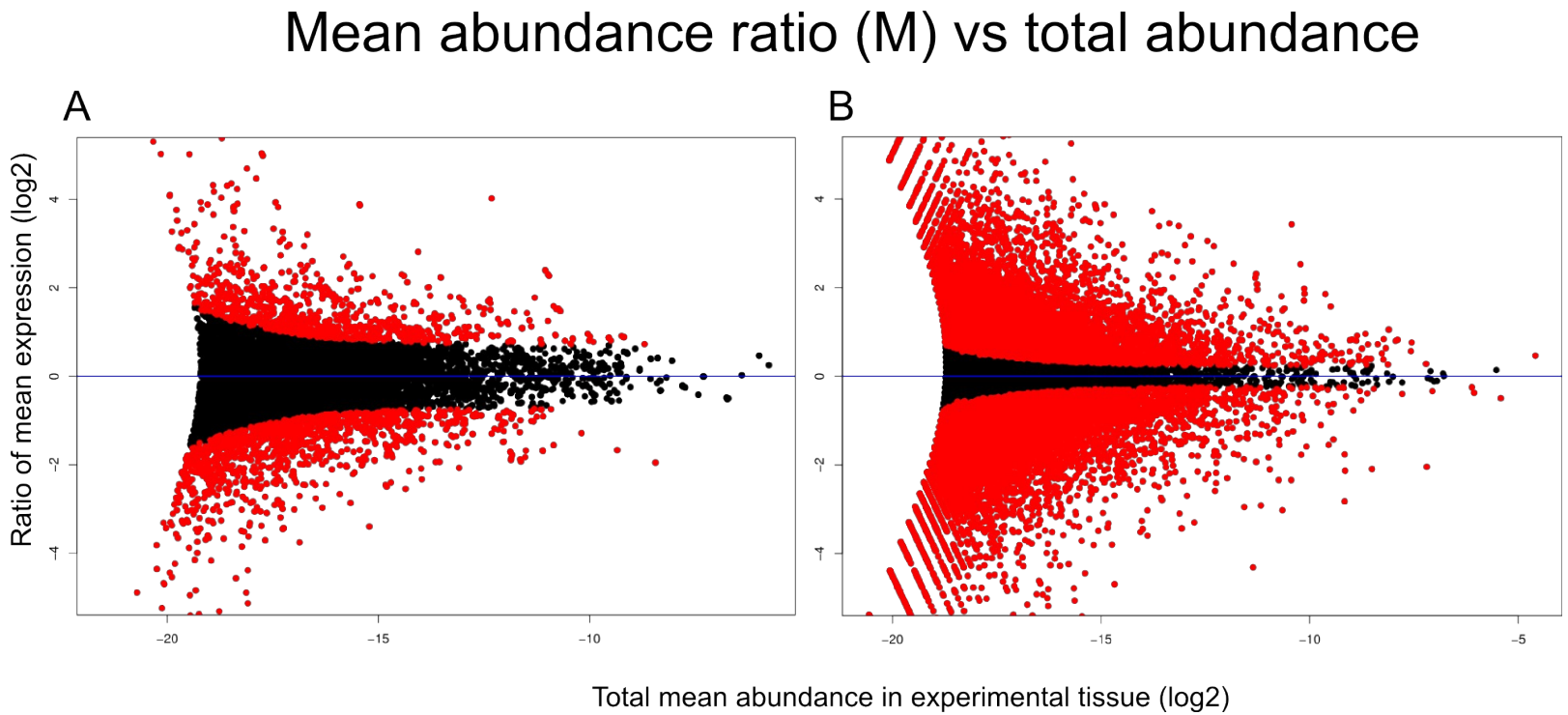


***Figure 12**: Plot of expression ratio of means from female vs male sample (log2) versus total abundance as sum of the two sample means (log2) with the two procedures of estimating significance: A) contig approach; B) tag approach. Both approaches had low count instances removed. Black spots shows comparisons judged non-significant and red spots are significant at an FDR of 1%. Note the higher dispersion of points in the tag approach. This results in part from two unique tags collapsing to the sage restriction site and in lesser part from two tags aligning to one contig.*

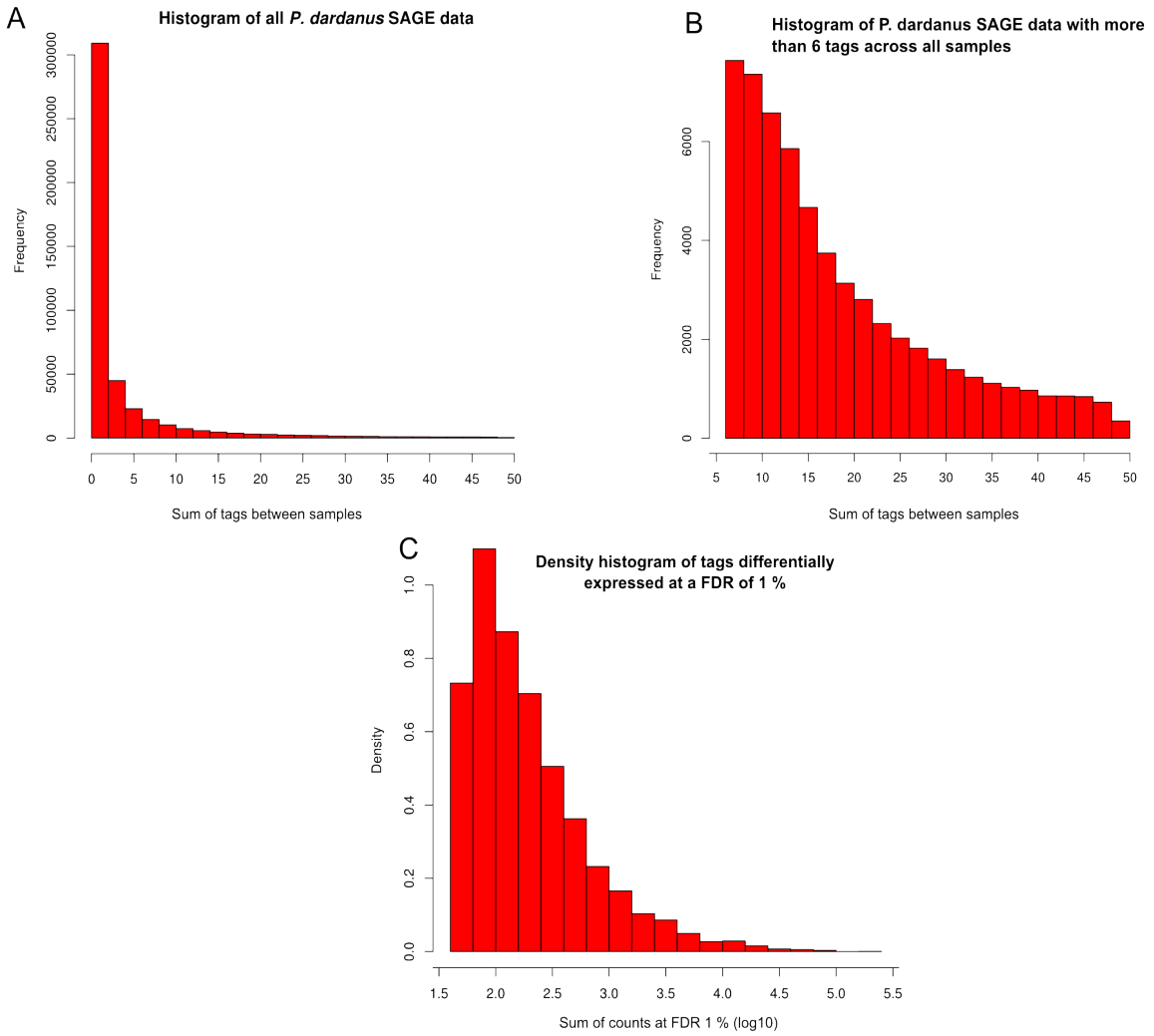


***Figure 13**: Expression level across both samples in tag counting approach. A) Histogram of all tags with coverage less than 50 counts; B) Histogram as tags with less than 6 counts are removed; C) Density histogram of tags judged significant at a 1 % FDR level.*

Table 7 Melanogenesis and pre-patterning candidate genes identified in P. dardanus and P. glaucs, P. xuthus, D. melanogaster, B. mori. Here the % pairwise identity of the P. glaucus copy to other species on the nucleotide (nt) and amino acid (aa) level is shown. Pd signifies P. dardanus, Px P. xuthus, Bm B. mori and Dm D. melanogaster. Dash (-) signifies gene not found in that species.

Gene name & synonyms	nt - Pd	nt - Px	nt - Bm	nt - Dm	aa - Pd	aa - Px	aa - Bm	aa - Dm
N-beta-alanyl-dopamine synthase (BAS; ebony)	-	84.4	64.4	57	-	91	61.3	44.7
dopa decarboxylase (DDC)	-	88	76.5	68	-	96.7	91	73.5
GTP cyclohydrolase IA (GTPCHI a; Punch)	-	83.3	80.9	73.1	-	97.5	93.3	80.2
phenylalanin hydroxylase (PAH; Henna)	-	84.4	70.9	62.7	-	95.4	86.5	69.7
tyrosine hydroxylase (TH; pale)	89	90.6	78.8	67.4	96.6	97.9	93	70.2
yellow	-	78.5	65.9	59.3	-	80.9	67.7	51.5
tan	-	82	64.3	54	-	88.6	65.1	41.8
laccase 2	-	89.6	77.8	75.7	-	95.9	88.9	75.6
black	-	87.4	75.1	62.7	-	95	84.6	61.6
purple	88.1	90.2	74.8	59.2	94	94.6	76.8	54.1
sepiapterin reductase	82.1	-	67.7	54.7	88.5	-	75.6	35.5
Pre-patterning genes candidates:								
enhancer of split m gamma (E(spl)mgamma)	-	-	54.8	53.2	-	-	44.7	33.7
rudimentary	-	-	-	-	-	-	-	-
fringe	86.9	-	68.9	55.5	89.4	-	87.2	52.8

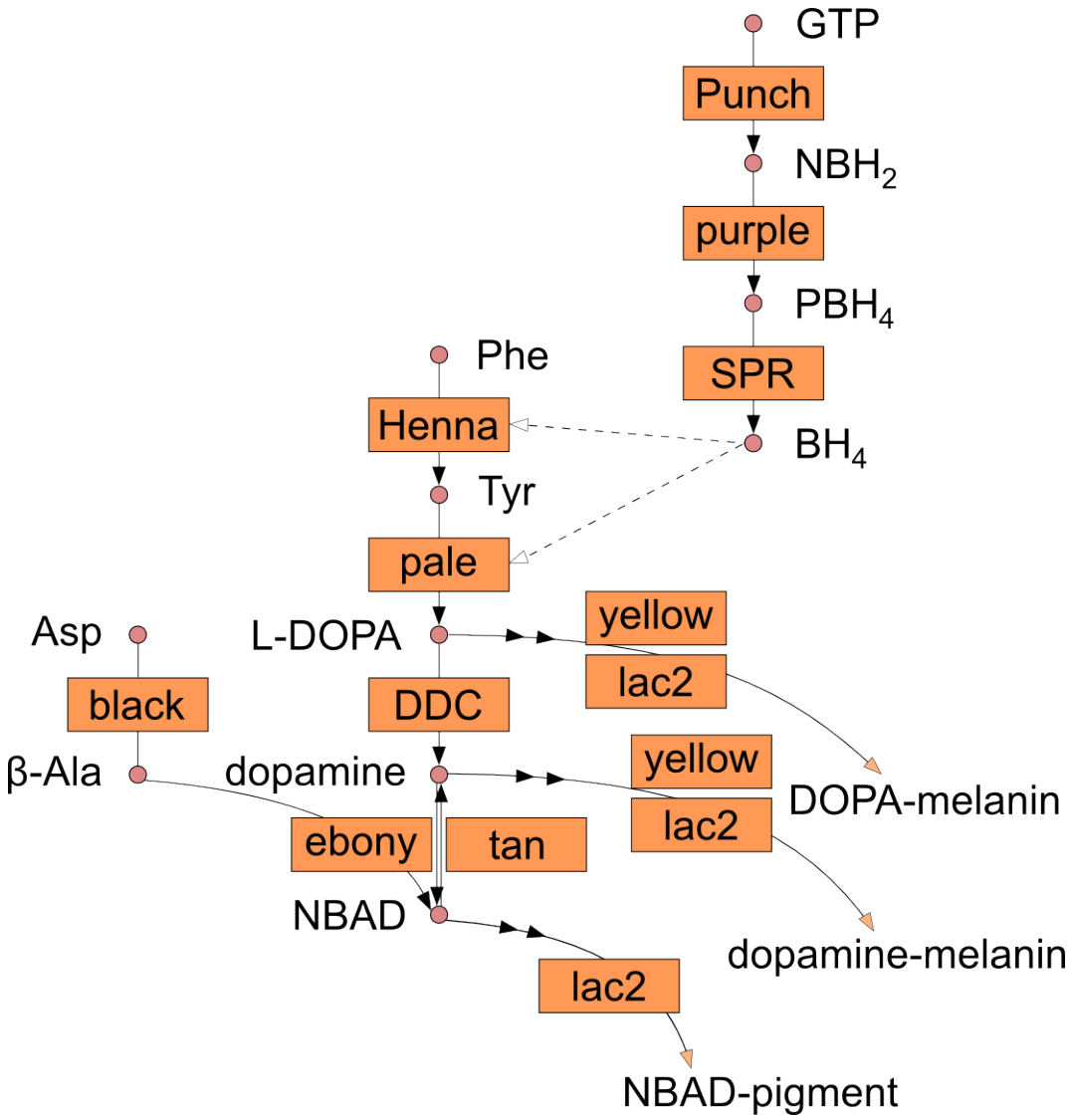


***Figure 14**: Melanogenesis pathway as reconstructed from the literature. See main text.*

Candidate genes

For use as candidates, I manually curated genes which are known to be members of the melanin biosynthesis pathway. Previous published work on Papilio glaucus and P. xuthus (Shirataki, Futahashi, and Fujiwara 2010; Futahashi, Banno, and Fujiwara 2010; Futahashi and Fujiwara 2006; Futahashi and Fujiwara 2005; Futahashi and Fujiwara 2007; Futahashi et al. 2008; K. Sato et al. 2008; van't Hof and Saccheri 2010; Wittkopp et al. 2003) allowed for the reconstruction of the hypothesized pathway (Figure 14). In general, the P. glaucus, being the better dataset, afforded more genes pointing towards the fact that for genes with low expressions, as those often involved in development, deeper transcriptome sequencing is required (Table 7 shows which genes were

found). Most genes showed high degrees of nucleotide conservation and in each case, conservation of annotated domains were used to ensure that the correct gene was identified. Because the 454 and the deep-SAGE data originated from the same individuals it would be possible to align the differentially-expressed SAGE data to identify if any of candidates were differentially expressed. By aligning the initial deep-SAGE reads, 835 reads aligned to genes found in P. dardanus or even P. glaucus showing that this approach could work. None of these reference genes ORFs aligned, however, with any of the tags or contigs shown to be differentially expressed. Due to lack of biological replicates, it is not known how heterozygosity has affected these results and also we cannot make an inference regarding non-curated genes. Using the InsectaCentral annotation, however, I was able to identify that the Laccase I protein had tags which were differentially expressed between samples; the two samples had a different tag. Unlike Laccase II, there is no evidence implicating Laccase I in melanogenesis.

It would be of use to clone these genes in P. dardanus and re-perform the alignments. Further, due to the small number of reads that did align, differential expression experiments for genes expressed in low levels seems to be problematic. Generally, the candidate gene approach is more robustly addressed by a real-time PCR methodology rather than a whole transcriptome scan.

Candidate loci

Table 8 Classification of tags assembling with H locus sequence

Type	Female specific	Male specific	Female biased	Male biased
Intergenic	1	2	9	14
Intron	2	2	1	5
Exon	1	0	0	0
Possibly erroneous	0	1	2	4
Possible UTR	0	0	1	0

Clarke and Sheppard defined the gene controlling appearance of the different mimetic forms in P. dardanus as the H-locus, and linkage mapping has identified a candidate region containing this locus (R. Clark et al. 2008). It has been partially sequenced using a BAC sequencing approach so I investigated how the differentially expressed tags further annotated the H locus contig. Even though the design of the Illumina experiment was not sufficiently robust to offer any conclusive information, this approach could shed light on what genomic region, especially repeats, influence the deep-SAGE results. The first step was to prepare a reference genomic sequence by assembling available BAC sequences to one contiguous contig spanning 339,759 kb. One of the BAC sequences contained an inverted region which was corrected in the final contig.

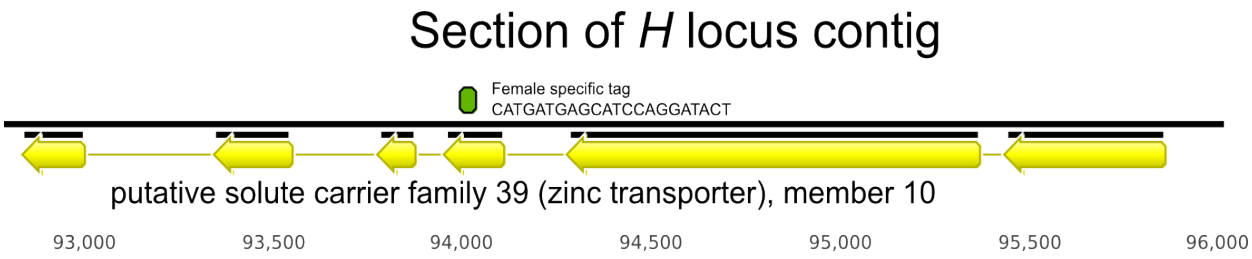


***Figure 15**: Section of the P. dardanus H locus sequence with a female-specific SAGE tag aligning with the exon of a zinc transporter protein.*

After assembling the tags found above to the H locus sequence, I categorized them in relation to their position with annotated ORFs (Table 8). As the data is cDNA derived, tags should align to regions of the genome transcribed as mRNA. Only one tag, a female specific one, was aligning to an exon sequence (Figure 15), making it a good candidate for a gene which may be differentially regulated between the melanic females and the non-mimetic males. In two instances, tags from conflicting groups overlapped. Even though the tags were collapsed using the fastx_collapser (which takes quality into account) it is likely that issues with heterozygosity should be addressed in the future as well as an improved experimental design.

Experimental design

When results such are these are viewed in the context of the transcriptome-wide annotations, such as the one stored in InsectaCentral, one could begin to build a picture of how the phenotype shapes gene expression of a particular tissue. In this particular experiment, the reference transcriptome is incomplete: it is derived from a single developmental stage with only two of the four potential genotypes (mimetic females, non-mimetic females, non-mimetic males and non-mimetic males carrying the mimetic allele; the latter two have identical phenotypes). Specifically, the non-mimetic females were not included and no genetic information was available to assign genotypes to the males. The experimental-design has been compromised in a number of additional ways. The Illumina experiment also lacked genotype information for the males, so one cannot be sure which variable caused the observed differences. During sample generation, multiple individuals were pooled in order to remove individual bias and be cost-effective. However, no barcoding procedure was used and therefore a single individual could (and probably has) skewed the expression levels of the entire sample. This can be effected in a number of ways. First, an infection or other external factor in one individual would cause a certain class of genes to be overexpressed in one sample.

This is particularly likely when the individuals originate from the field and then stressed by being raised in laboratory conditions using decaying plant material (Rod Mahon, CSIRO-Entomology Australia, pers. communication). Second, expression studies, digital transcriptomic or in-situ hybridizations, require accurate staging of the individual samples. The use of morphological markers has been controversial (Reed, P. H. Chen, and Nijhout 2007). One can use a reference gene (via estimating expression levels and/or patterns) but no suitable marker has been developed for this species and tissue combination. In such cases, it would be prudent to prepare a large number of individual samples and keep sequencing costs low via the use of barcodes. Statistical power gained from the use of multiple samples should out-weigh the decrease of sequence coverage. Further, use of barcodes, assuming they have similar GC content, would negate the need for any technical replicates even though current research shows that, unlike microarrays, the Illumina platform shows no bias between sequencing runs or equipment (Marioni et al. 2008).

Conclusion

This deep-SAGE experiment should be repeated with the a larger number of samples derived from single individuals. A suitable lab-colony, not showing any signs of infection, would be the most robust solution. A further complication exists with restriction digest based methods using pooled samples: mutation of the restriction site in one individual of a P. dardanus sample would cause the perceived expression counts to drop by 1/7. Due to the high heterozygosity present in Lepidoptera, it is likely that a number of the sex-biased results is due to such an effect. Likewise, genes which are differentially expressed may not be detectable. A solution to this issue would be to have access to the expression levels of individuals rather than pools and to perform two experiments with two different restriction enzymes. The curation of the pathway is a significant contribution and should be expanded with full-length sequencing. A wet-lab approach using real time PCR should explore the differential expression of the candidate genes presented here. Careful generation of material for samples would require a steady supply of appropriately genotyped individuals, a resource currently lacking in this non-model species. Overall, however, the tag method presented here has the potential of dissecting the developmental processes with higher throughput and is gene agnostic. It will, however, require more samples in order to identify tags which are significantly differentially expressed within a statistical framework.

Overall chapter synthesis

Can a transcriptome reference sequence function as an anchor of -omic data when studying a biological phenomenon? Can it thus substitute for a genome-based reference when this is

unavailable and too expensive to produce? The above case studies on species lacking a genome sequence, answer these questions with a 'yes' within limitations. They can also provide further insights via their commonality. A more pressing question is, however, 'how can we best design and make use of transcriptome-based studies?' It is, first, obvious that novel routes are now open in both resource-model and non-model species research. This ability is reliant on not only the NGS technologies or computational resources but also novel computational approaches which allow to process such data. Many of these approaches have been developed thanks to genome sequencing projects. Nonetheless, for this chapter new methods had to be devised or used in order to allow a transcriptome sequence to be used as a reference to study a biological phenomenon. The field of transcriptomics without a genome reference is novel and requires that improvements are made.

Technological innovations

Sequence information is no longer a limitation: the operational cost has shifted. Sequencing of entire mitochondria (McComish et al. 2010), direct sequencing of RNA (Ozsolak et al. 2009) or Next-Gen sequencing of PCR products pools is changing the way laboratories operate. Technologies are evolving rapidly offering new methods such as the deep-SAGE described here. This technique involves sequencing a few bases downstream of a restriction site. New Illumina technologies allow us to now increase the number of bases to 100 bp allowing for better characterization of species with high levels of heterozygosity. Generating SNP information was, until recently, laborious. The est2assembly platform produces it as a by-product of transcriptome sequencing. It remains to be seen if the Illumina Bead-station is the most efficient protocol for NGS genotyping of individuals. It is conceivable to utilize the deep-SAGE to acquire SNPs linked to a restriction site. This is similar to another protocol, RAD-TAG (Baird et al. 2008), but would allow us to both measure expression levels and acquire SNP information linked to coding sequence. New, more efficient barcoding and sequencing protocols can increase the number of individuals which can be pooled since this technique relies on sequencing a reduced representation of the genome or transcriptome. Further, new technologies such as the Ion Torrent may allow for research groups to use a benchtop sequencer for NGS genotyping and thus not rely on sequencing centers. The technological advances must however be used within a well designed project.

New operational paradigms

Utilizing public data to generate candidate genes or phylogenetic trees is not new, but the availability of these data is. For phylogenetics, NGS-based transcriptome sequencing produces such

a wealth of data that phylogenetic information will be produced inevitably. It is, however, not the norm to make these data public, which is why InsectaCentral was built. This platform allows researchers to store and mine data for a use that may be irrelevant to the original project that produced these data. The stored data are, however, not curated. Phylogenetics is particularly prone to the effect of erroneous bases. A phylogenetic framework is an essential background for functional biology, for example studies attempting to understand the evolution of protein families. Students of protein family evolution can, perhaps for the first time, hope to study their favourite family within a rich phylogenetic framework. However, the first case study shows us how we can approach phylogenetic research systematically and point towards future directions. Future workers in Phylogenetics/Phylogenomics must overhaul their approach in multiple areas. First, current evolutionary models are sufficiently robust to address these new datasets but computational approaches utilizing them in an efficient manner may not be. Second, as molecular evolutionary biologists from this and other fields utilize the comparative approach, they have to address the problems of saturation and composition bias. New computational hardware technologies and High Performance Computing can provide the results rapidly, a fraction of the time required for the construction of an average phylogenetic ML tree. A solution to composition bias can be based on pairwise comparison of sequences and the use of simple statistical tests. It would make sense to include these tests within a phylogenetic software. One issue that has not been taken up widely is the utility of TreeBase (Piel, Donoghue, and Sanderson 2000). Supporting and improving a resource which can warehouse trees and original data used in publications of phylogeneticists is important in order to ensure replicability and allow for better comparisons. InsectCentral should be improved to interface with it.

We would need a similar framework for 'digital transcriptomics' or candidate-gene hunting. It is common for genome database to house transcriptomic data such as microarray experiments (Duan et al. 2010). One issue with such an enterprise is that experiments would have to conform to certain specifications, including the relevant Minimum Information Criteria (cf. Stoeckert, Causton, and Ball 2002). Currently there are no guidelines for Illumina-based surveys of gene expression yet it seems that Illumina-based approaches are more reproducable, more accurate and require less pre-processing (Mark Blaxter, pers. communication; Ruzanov and D. L Riddle 2010; c.f. M. D Robinson and Oshlack 2010). Pooling of individuals is not recommended. Further longer reads would be most beneficial for outbred species as alignment programs require a sufficiently long sequence to act as a 'seed', i.e. be an exact match to the reference, prior any alignment extension. The exact implication of polymorphism-richness in the sample libraries remains to be seen but

improvements in the alignment software could include the ability to allow for degeneracy. This would not solve, however, the problem of insertion/deletion polymorphisms (indels). An important take-home message is, however, that an informed project design is important. Like in microarray experiments, the background noise must be minimized. Unlike, microarrays, however, we have the option of manually checking alignments. Raw sequence data are more informative than the fluorescence levels of microarrays. Replicates are still needed, and allowing for redundancy would allow for a failed run (e.g. M. sexta Control3). Finally, the utility of having a transcriptome reference of high quality can be easily understood with the deep-SAGE studies. A curated reference would remove spurious results that can compromise the statistics depending how this is approached. If tags are not grouped by gene but investigated independently, we will be able to identify alternative splicing or differential expression of isoforms. Integrating this knowledge on the gene level would only be possible if we had access to a curated transcriptome. With this approach, our coverage per gene would, however, be lower. If tags are grouped by gene, it might be more accurate to measure expression levels of entire mRNA transcripts but would comprimise our ability to detect alternative splicing, as the one expected from sexual dimorphism (e.g. P. dardanus dataset). Other problems are specific when grouping by gene. Two contigs may be from the same gene because indels or polymorphism prevented them from collapsing. The 3' UTR is particularly prone to that effect. A further effect of this phenomenon is that tag count belonging to different parts of the same mRNA transcript may not be grouped for certain genes. One should keep in mind, however, that these candidate predictions are only predictions and must be verified by a wet-lab approach. But is it possible therefore to conduct differential expression experiments and generate candidates without a genome reference? Yes, but an effort for curating the transcriptome must be invested. With the inexpensiveness of Illumina sequencing this may become easier and it would be important to invest in further developing est2assembly to accept RNA-seq data.

Author contributions

For each of the studies presented in this chapter, I was only responsible for designing, conducting and analyzing the bioinformatic component and drafting any manuscript(s) unless otherwise stated. For the reference transcriptomes, complementary DNA libraries were contributed by a number of people as mentioned in the Materials and Methods section or the est2assembly chapter of this thesis. Briefly, the Manduca deep-Sage project was initially designed by Dr. Yannick Pauchet (University of Exeter) and Marian Thomson (University of Edinburgh) and the approach to analyse data was designed by myself with assistance on GLM modelling by Dr. David Hodgson (University of Exeter). The deep-SAGE on P. dardanus and the RNA-Seq on C. tremulae experiments were

designed by Prof. Richard ffrench-Constant and conducted by Dr. Iva Fukova with the analysis designed and performed by myself. The SNP project design was undertaken with Dr Jon Slater (University of Sheffield) and Prof. Tom Tregenza (University of Exeter), the two stakeholders of the G. campestris project; bioinformatic analysis was performed by myself. The RP project was designed and phylogenetic trees were generated and analysed by myself but curation and protein alignments of insect RPs were initially generated by Ms. Victoria Renders (University of Exeter) as part of her BSc thesis under my supervision. Dr Lars Jermiin provided crucial assistance in the composition heterogeneity component. The curation of the P. dardanus BAC and the melanogenesis pathway was conducted by myself with assistance from Dr. Iva Fuková (University of Exeter).

Overall discussion

The aim of Large Scale (LS) experiments is hypothesis generation (Collins et al. 2003). A typical LS experiment consists of the following phases: model system selection; (optionally) resource generation; production of experimental data; analysis/candidate determination and hypothesis generation. The subsequent stage is similar to a traditional hypothesis-driven research: validation; model refinement and hypothesis reformulation. Coupled with a critical approach and an unbiased methodology, the above is a summary of the scientific method. The sole difference between LS experiments and hypothesis-driven research is that LS experiments set out with no explicit hypothesis. This first stage is an inherent property of all -omic fields and focus of this thesis. Specifically, this thesis addresses the scenario that a question-model is also resource poor but sufficient funds exist to elevate this condition. Only sequence resources are considered here but there is no reason why not to apply the findings to non-sequence resources.

Model system selection and resources

Selection of a model system is a biologist's prerogative. An example from the Lepidoptera (butterflies and moths) model system for colour pattern variation and evolution of mimicry has been laid out in the first chapter (Beldade et al. 2007). Among those organisms outside established model systems, butterflies offer exceptional opportunities for multidisciplinary research on the processes generating and maintaining variation in ecologically relevant traits. In that Chapter, my co-authors and I highlighted research on wing colour pattern variation in two groups of Nymphalid butterflies, the African species *Bicyclus anynana* (subfamily Satyrinae) and the South American genus *Heliconius* (subfamily Heliconiinae), which are emerging as important systems for studying the nature and origins of functional divergence (Beldade et al. 2005; Joron et al. 2006). At the time of writing (i.e. 2007), we predicted that growing genomic resources (e.g. genomic and cDNA libraries, dense genetic maps, high-density gene arrays, and genetic transformation techniques) are extending current gene mapping and expression profiling analysis. These would enable the next generation of research questions linking genes, development, form, and fitness. Since 2007, many of the above resources have been produced as well as a genome for the *Heliconius melpomene* butterfly with an N50 of more than 150 Kb showing that our predictions have been correct (Baxter et al. 2010; A. Monteiro & Prudic 2010). In this thesis, I have started the focus with Lepidoptera, proceeded to be inclusive within all insects. The work of this thesis aims, however, to be species-neutral and therefore the above taxa were only used as development datasets or case studies.

Producing a reference transcriptome

Overview

The est2assembly program (http://est2assembly.googlecode.com) is responsible for producing high quality transcriptomic assemblies from Sanger or NGS raw data (Alexie Papanicolaou et al. 2009). Being able to also analyze and disseminate the results, at the time of writing it is the only platform of its kind. The analysis component was written in Perl and offered plugins for the only two assemblers known to be capable of analyzing transcriptomic data: MIRA (Chevreux et al. 2004) and Newbler (454 Life Sciences). For dissemination, it utilized the Bio::SeqFeature::Store database schema to drive the GBrowse software and also used Chado for data-warehousing (Stein et al. 2002; C. J. Mungall & Emmert 2007). The program utilizes a number of other software which are golden-standards in their field: SSAHA2 (Ning et al. 2001), BLAST , prot4EST (Wasmuth & Blaxter 2004), InterProScan (Zdobnov & Apweiler 2001) and annot8r (Schmid & Blaxter 2008). The advantage of est2assembly is that it is not written as a single software but as a platform with modular components. This allows current and future developers to expand and shape it according to novel software that is produced. For example, after est2assembly was published, I developed a digital transcriptomics module which utilized BowTie (Langmead et al. 2009) and the R statistic package (Team 2009) in order to be able to process data such as those present in the case studies chapter. Likewise, only minor changes had to be made in order to support the latest version of MIRA (version 3). The modular nature has added advantages: new software can be integrated relatively easily and swapped. Users can select how to transverse the pipeline, i.e. which components to make use of and how. This modular architecture is similar to the one from CABOG, also known as Celera Assembler or wgs-assembler (Jason R Miller et al. 2008). CABOG is the golden-standard in genome assembly and, like est2assembly, is also a collection of 50-odd scripts tied together with a pipeline.

Shortcomings, solutions and future directions

As it relies on 3rd party software, a number of problems do exist. The platform was written with high-throughput users in mind, such as a sequencing center. Accepting raw output directly, it first pre-processes the data with users having a high control of how it is processed. The processing is, however, optimized for Sanger or 454 data. Indeed, the platform does not support the Illumina platform because at the time of programming the standard read size for Illumina was only 35 bp meaning that it would be of no use to transcriptome assemblies. With read size having now expanded to 110 bp and increasing to more than 200 bp with the new Illumina Hi-Seq machines, it

would be important for est2assembly to support this technology. There are currently plans to integrate this so called RNA-seq technology as part of an est2assembly module possibly using the new Columbus module of the Velvet assembler (Zerbino & Birney 2008). Columbus is, however, an experimental module and also the memory requirements would be prohibitive to the ordinary user. Programs such as those might help with an additional deficit of est2assembly: handling of highly polymorphic data. Computationally it is not straightforward to distinguish whether two similar but not identical sequences are paralogues or just alleles of the same locus. The current approach aims to generate a single reference and therefore works by discarding degenerate contigs. This has the side-effect of potentially losing isoforms which are highly similar. The trimming procedure, like all est2assembly routines, is however customizable. After a subsequent automated annotation step, there are two possible paths the user can follow before using the transcriptome for a downstream application. One is to utilize the automatically generated reference contigs, keeping in mind that not all SNPs have been accounted for and that highly polymorphic regions, such as the UnTranslated Regions (UTR), might be present in redundant but diverged copies or even in alternative forms (Mangone et al. 2010). The alternative path is to undertake a manual curation effort. The choice largely depends on the nature of the downstream application and the human resources available. In the phylogenetics case study, the input data had to be of the highest quality and thus manual curation was used. In the deep-SAGE applications – based on Illumina sequences adjacent to a restriction site on mRNA transcripts – there was more concern whether short reads would align to a reference sequence derived from a library with a high number of haplotypes (I'm not currently aware of any short-read alignment software which can accept degenerate base information as they were all desgined for the processing of raw sequence data). Further, the aim of the application was to provide a global view of transcription differences and manual curation of an entire transcriptome is a time-consuming process. One alternative is to utilize the relatively novel RNA-seq technology (Illumina sequences from the entire mRNA transcript) to saturate coverage allowing for discovery of alternative isoforms with both the Open Reading Frame (ORF) and the UTR. This approaches has yet to be tested on a species without a genome sequence but a project on the *Heliconius erato* species is underway.

In libraries derived from multiple outbred individuals, a number of SNPs are often present in the coding sequence but rarely cause problems in the assembly. Even limited alternative splicing can be handled by the modern assemblers used by est2assembly. The major issue with most question-model transcriptome projects is that the scientists involved wish to accomplish goals with conflicting requirements: production of a high quality reference sequence; saturation of gene

finding by using multiple expression profiles and polymorphism detection. Unless inbred samples are used, the library will be populated by a large number of haplotypes. While coding sequence SNPs are sufficiently rare and dispersed, UTR divergence can be significantly high to prevent contigs from forming overlaps and almost certainly de-Bruijin graphs. If multiple libraries were available, the correct approach would be to use a subset of the data to build a reference and use the remaining data to produce mapping assemblies and SNP information. An automated method could be deployed in the future. The UTR issue is one of the most misunderstood issues about transcriptome assemblies. When the 3' UTR is large, contig inflation in inevitable, as is depicted in the Venn diagram of Chapter 1. A number of these contigs will indeed be coding, but many will just be long 3' UTR regions with multiple polymorphisms to collapse to a single non-redundant contig. Further, research using the new Helicos platform (direct RNA sequencing rather than cDNA) also confirms high heterogeneity of the 3' UTR (Ozsolak et al. 2009).The problem is exaggerated downstream software, such as prot4EST, expects that all contigs are coding and therefore attempts to produce an ORF for each. Without making some ad-hoc decision on how to decide between paralogues and alleles it is not possible to resolve the issue with existing algorithms. One idea would be to have a program which is capable of making a decision based on ancillary data, such as GC content, presence of stop codons or codon usage (or lack of it) and others. Currently no such software exists. It would be of interest to write a module in est2assembly which can perform such a task semi-automatically and then also try to extend the contigs. Contig inflation can be addressed by a pairwise alignment approach in order to identify redundant datasets. The trim_assembly step in est2assembly uses a global alignment approach. An alternative approach, utilized by the University of Edinburgh Sequencing Service, is to make use of the Minimus2 program from the AMOS package (Sommer et al. 2007). This software utilizes an overlap graph but was built for genomic data and the overlap based approach is unlikely to be appropriate for the relatively small transcriptome contigs; further, the low diversity in a coding region coupled with high diversity at the UTR ends would have the signature of a misassembled genomic contig. Indeed, the lack of software programmed for transcriptomic data is the major limitation of transcriptome sequencing: as mentioned in the introduction, in the bioinformatic field transcriptome sequencing is undertaken after whole genome sequencing and assembly. Most existing assemblers were primarily designed for bacterial genomic data and then subsequently modified for eukaryotes. Some, like MIRA, were further modified for cDNA datasets but all come short of producing good assemblies from libraries with outbred material. The main reason is that it is difficult to optimize an assembler for what is essentially a signal-to-noise problem: what is considered noise at a genome sequence (short contigs) may well be a signal in a transcriptome (short gene). Furthermore, assemblers using de-Brujiin

graphs – and to a lesser extend, overlap graphs – rely on distribution of oligomers to determine repeat status of a sequence (J. R Miller et al. 2010). Transcriptomic data, even when normalized, do not have a uniform distribution confounding, thus, such assemblers. The more traditional assemblers using a greedy extension, such as phrap (Green 1996), do not build an initial overlap graph and contig extension is a one way path. This results in the nature of contigs being influenced by the order of clustering and if a misassembly occurs, this error propagates through the rest of the assembly process. Currently the MIRA3 assembler seems to sufficient for most needs. A major innovation of est2assembly was the ability to use multiple assemblers, explore the parameter space and provide an objective benchmark based on coverage of a reference transcriptome. Therefore, once available, an assembler designed specifically for EST dataset could be included it in est2assembly as another plugin. Until then, it is best to generate reference transcriptomes from as few chromosomes as possible and for downstream applications requiring high quality reference, manual curation is still necessary.

Disseminating a reference transcriptome via a robust infrastructure

Overview of software

With a reference and annotation now available, the data must be integrated, made available for mining and perhaps for curation, i.e. editing by a human. Integration is accomplished by storing them into a common data warehouse. Data mining/curation requires a specialized user interface (UI). This UI is then responsible for presenting data in a specific and structured way. It can be part of a generic data-mining software (such as the ubiquitous Gbrowse) or specific to a database instance. The Drupal UI is built using my custom Drupal modules published under the genes4all project (http://drupal.org/project/genes4all). They make use of my gmod-dbsf library (http://drupal.org/project/gmod_dbsf). This generic library was used to build another UI for deploying bioinformatic software servers such as the (also ubiquitous) BLAST. All UIs require an efficient method for storing and retrieving data from a source. The most structured approach to store data is in the form of a relational database. Its outline and structure is called a schema and defines how secure and fast is the manipulation of data objects: normalization decreases both the chance for data loss and speed. In genomics, the commonest database which handles generic data and is highly normalized is GMOD's Chado, initially a FlyBase project. Such a schema is valuable as a data-warehouse but is too slow for supporting UIs. For example, FlyBase pre-computes much of the data available on their web-pages and stores them as XML. Other techniques within the ARGOS package, such as distributed and load-balancing servers, ensure real-time responses (Gilbert, pers.

communication). Further, editing or curating data requires provisions for data security. The approach undertaken in the InsectaCentral work is utilizing all three steps. GMOD's chado acts as a data warehouse. BioPerl's feature store acts as a generic software schema. The Drupal database is used for bidirectional UI such as curation but also serving cachable web content. Further optimizations are delivered via materialized views. A complex query is one which demands data from multiple sources of information. These are computationally intensive and thus time consuming. In InsectaCentral, these queries are fixed and stored as virtual database tables (views). The query is then pre-computed and stored in the warehouse, materializing thus the virtual table. Data updates require re-computation of these materialized views and therefore they are only suitable for data which is not updated often. This technique is coupled with the ability of gmod-dbsf to use “caching”. Gene pages or searches can still take a non-trivial time to pull data out of the warehouse and produce HTML code. Caching stores an HTML code for a configurable amount of time (e.g. a week in the InsectaCentral implementation) so that identical searches execute at a fraction of the originals. This allows databases such as InsectaCentral to host and serve millions of data points.

Overview of InsectaCentral implementation

Even though the genes4all software described above is species-neutral, the InsectaCentral implementation produced in this thesis is populated with insect transcriptomes. Other attempts to build an Insect-wide database have not come to fruition (Chris Elsik, pers. communication). Current plans of the 'ArthroBase' are focused only on species with sequenced genomes because there is, apparently, no perceived need to expand to resource-poor model species. InsectaCentral's mission is to allow the resource-poor model species researchers working without database funding to make use of their own and the community's data in a streamlined, efficient and standardized fashion. The NGS data derived from collaborators are complemented with those acquired from NCBI's dbEST, Short Read Archive and GenBank (Boguski et al. 1993). A feature of InsectaCentral is the deep annotation using BLAST similarity analysis to many databases, electronic inference annotations from large ontology sequence databases and InterProScan domains. This collection of resources were first introduced to Lepidopterists in ButterflyBase and was considered successful. If a relatively small project such as ButterflyBase produced a Faculty of 1000 citation and a double digit citation number in its short life span, then an Insect-wide database is clearly needed by the community. The specifications of such a resource must be conceptually different from those of a genome database. For example, because new transcriptomic data for any one species may be produced, gene models are expected to change. Using the unique identifiers of est2assembly, a

curator can re-compute assemblies and disseminate the information without invalidating previous data. Further, the structure of est2assembly and the Drupal modules allows an efficient yet permanent storage of the data. It currently hosts 12,800,018 ESTs forming 1,518,114 contigs in a data-warehouse of 100 Gb. This is complemented by a total of 189 Gb of Feature::Store databases used to drive the GBrowse software. In comparison, FlyBase hosts an order of magnitude less genes and ENSEMBL hosted 80 Gb of data in 2008 (Stalker et al. 2004).

Shortcomings and impact

The obvious shortcoming of InsectaCentral is the lack of support for genomic data. This thesis, however, is focused on the deployment and utility of transcriptomic data. Indeed, the support for genomic data is trivial to implement: a number of groups have been producing relevant tools and procedures for years. A less obvious but most important deficit is the lack of curated sequence. A number of models have been developed by the genomics community: a dedicated curation team (e.g. FlyBase (Wilson et al. 2007)), volunteers (e.g. VectorBase (Lawson et al. 2009)) and, recently, a Community Annotation System (CAS, e.g. in AphidBase (Gauthier et al. 2007)). Unfortunately, as there is no dedicated man-power/funding, manual curation of the datasets is not possible. Likewise, without a dedicated curator, a CAS cannot be implemented. The latter is, however, of particular interest considering the volume and diversity of the data. Further, with no reference genome, curation would not be focused on intron/exon structure but ORF/UTR, isoform and alleles. Orthologue and paralogue identification would also be important. Orthology is a third shortcoming of any transcriptome based dataset. Due to the fact we do not have a complete sample of the transcriptome, distinction of paralogues from orthologues is a non-trivial issue. A number of approaches, build for species with whole genome projects, are available such as inParanoid (O'Brien et al. 2005), TribeMCL (Enright et al. 2002) and others. A number of groups have attempted to utilize them for ESTs (e.g. James Wasmuth, personal communication and PhD thesis; COMPARA pipeline of ENSEMBL) but they can only offer predictions based on arbitrary cut-offs. My early investigations have concluded that an approach which would take into account the mode of evolution of a particular gene family and the phylogenetic structure of the species involved would be challenging to generalize. Recently, I identified the importance of presence of composition bias for the same problem (see phylogenetic case study in Chapter 6). A procedure straightforward to implement would be to use a reference genome. For each set of taxa, one reference genome would be used to anchor predicted ORFs and thus predict sets of paralogues. The quality of these predictions would be biased by the quality of the reference and the distance to each taxon. Further, the most important issue, distinguishing between alleles/isoforms and paralogues,

would not be solved as we know that protein family composition is dynamic (Hahn et al. 2007). InsectaCentral's automated annotation already provides a standardized solution to the problem of retrieving sets of genes using orthology by similarity, eliminating thus the intensive step of selecting reference genomes. In any of these approaches, a most important complication is, however, that orthology by sequence similary does not imply orthology by function and vice-versa. Most scientists are interested in the latter, yet current bioinformatic approaches are focused in the former. Integration of experimental data is therefore of importance. This would include data presented in Chapter 6 and biochemical experiments. Indeed, a community database of the scope of InsectaCentral ought to address any data-types produced by the community. This is indeed the modern function of databases: not only to provide storage for data but to integrate them and provide a platform the community to organise. In a recent paper by Terenius, Papanicolaou et al (Terenius et al. 2010) we used InsectaCentral to compile unpublished data from RNAi experiments and investigate on why RNAi does not work in Lepidoptera. The UI was prepared using gmod-dbsf and minimum information criteria as designed by the working group. Existing work provided by the MIARE group was utilized (Minimum Information About an RNAi Experiment; http://miare.sourceforge.net) to develop controlled vocabularies. This resulted in the first incarnation of the genes4all_experiment module which can be adapted for databasing other kinds of experiments. As a pilot on the ability to populate a community database with a research community's participation, it was highly successful: the paper was authored by 70 scientists from 42 institutions in 21 countries (Terenius et al. 2010). The genes4all and InsectaCentral serve three functions already: a transcriptome resource for researchers working on insects, a community platform in which they can organize and a software package they can use to database their pre-publication data.

Utilizing a reference transcriptome for evolutionary biology

It is important to note that the work presented in this thesis is being used to support research in evolutionary biology. This is accomplished either indirectly (e.g. ButterflyBase citations) or directly (e.g. Chapter 6, co-authored manuscripts which appear in the Appendix or others in preparation). In this work, this has been accomplished via three avenues: a reference transcriptome to aid genomic sequence annotation (e.g. co-authored paper in appendix) or phylogenetics, polymorphic marker identification and investigation of transcriptional differences between sample treatments. Species phylogenies are useful also beyond basic systematic questions. A number of studies have or are being published on multi-locus phylogenies (Wiegmann et al. 2009; Longhorn et al. 2010). Further, a number of researchers publish on the evolution of specific gene families (Vieira et al. 2007;

Oakeshott et al. 2010). Having access to an accurate species phylogeny (including branch lengths) can help us explore how protein families evolve by comparing evolutionary rates of a specific members of a family, with the average evolutionary rate obtained from multiple loci. Indeed, a graph of branch lengths for each locus could be plotted and compared with the branch length of family members. Rapid evolution is one of the signals for functional diversification even though a small number of changes after duplication may also result in a novel function (J Zhang 2003). This rapid evolution is often site-specific (Y. Wang & Gu 2001) and therefore evolutionary rates must use a branch-site model (J. Zhang et al. 2005) in order to detect them. These so called "phylogenomic" approaches can be a powerful tool for prediction of function if used correctly (Sjolander 2004). It would be important to have a large number of species in order to identify the evolutionary event which triggered functional diversification. Ultimately, it is not possible to rely the future existance of large number of finished reference genomes as even the 12 Drosophila genomes are not considered as finished (Hahn et al. 2007). But as I showed here, it is possible to generate multi-species, multi-locus phylogenies used transcriptome data. Further, I utilized matched-pair tests (i.e. pairwise comparisons) of symmetry to identify any composition heterogeneity that may be present (Jermiin et al. 2008). Such heterogeneity is the result of a violation of one of the assumptions the General Time Reversible (GTR) model. As most evolutionary models used in phylogenetics are special cases of the GTR, one can only change to an even more general (and therefore parameter-rich) model. Over-parameterization is likely and most researchers prefer to account for heterogeneity instead. It was known from previous studies (Savard et al. 2006) that 3rd codon sites exhibited this phenomenon providing a improbably tree (holometabolous insects were not monophyletic). I detected, however, that 1st codon sites exhibited this phenomenon as well. Accounting for composition heterogeneity is currently only possible by recoding the third codon sites to the degenerate nucleotides R and Y (where this change is synonymous). For first codon sites, as recoding is rarely synonymous, the only solution is to remove the affected sites ("strip the column"). With these corrections a tree which does not violate the GTR model is possible. Even though the tree topology was not affected, branch lengths were. As mentioned, accurate branch length are important in determining rates of evolution of gene families. The disadvantage of accounting for heterogeneity is not over-parameterization but decrease of the available signal (i.e. number of informative bases in alignment). As a result, the work presented requires an increased number of genes in order to be more useful. The methodology, however, presented here shows that additional genes or species are straightforward to add. The limiting step is the manpower to curate and produce a high quality reference transcriptome.

The advantage of a high quality reference transcriptome is that this reference sequence can be used to support functional biology in other ways. The deep-SAGE protocol (case study of *Manduca sexta* and *Papilio dardanus*) was a powerful approach of generating candidates when few existed. In the M. sexta dataset, the known candidates (CYP450 enzymes) were identified. It could be assumed that these enzymes appeared by chance because the insects were stressed but only a small proportion of identified P450s were judged significant (4 of the 22 identified in the reference). This fits with the data presented by Stevens et al (Stevens et al. 2000) which shows that there is an array of P450s whose expression profiles can be surprisingly specific to certain xenobiotics. These candidates were later verified by collaborators (Dr. Y. Pauchet, Univ. of Exeter) using Real Time PCR. Bioinformatics, however, can only provide candidates, it can formulate hypothesis but these experiments cannot provide proof by themselves. A detailed biochemical study is required. Further, the results of the biochemical study can then assist in redesigning the transcriptomic experiment to be more specific. In some cases, like in this experiment, the bioinformatician(s) can see evidence that more samples are needed. Alternatively, as in *P. dardanus* experiment, experimental design can compromise the candidate lists especially when there is no previous body of work that could indicate some positive controls. It is important to note, however, that there has been no cause to doubt the replicability of sequence-based expression profiling (Ruzanov & Riddle 2010). This is further assisted by the fact that one can explore and verify the alignments. This allows us to extend the number samples after initial work has been carried out. Indeed, the conclusion is that more samples need to be run for the deep-SAGE case-studies but as a pilot experiment the *M. sexta* dataset has fulfilled its intented function. It would be of interest to the relevant researchers to improve the curation of the transcriptome and then re-run the pipeline in order to improve the statistics.

Indeed, a high quality reference transcriptome is generally important. For example, in the annotation of an eventual genome project (the *M. sexta* genome project is being planned). Throughout the project, identification of the number of genes available in the genome assembly can assist with estimating progress. This is a valuable information, supplementary to the commonly used N50 statistics (N50 index is the minimum number of contigs which can account for half of the assembly and N50 size is the size of the smallest of those contigs). Even though not commercially available during this thesis, the development of new Illumina RNA-seq protocols allow us to produce millions of sequences (of ca 200 bp in length or 200 - 400 bp if a paired-end approach is used) from entire mRNA transcripts with a small financial investment (ca $1,000 – $2,000 USD). It is possible that we will have high coverage transcriptomes well before the relevant genomes become available. This high-coverage will also reduce the manpower required for curation.

These experiments would have been challenging and costly to perform in resource-poor species without the use of NGS technologies. In this work, they did not require a reference genome, They would not have been useful without the bioinformatic protocols presented in this thesis. Together with a good project design, these have the power to transform evolutionary biology, at least the functional component, i.e. Ecological and Evolutionary Functional Genomics (EEFG). We're at the beginning of this new paradigm and the years ahead will be most interesting.

Overall impact and future work

In this work, I co-authored a perspectives paper (Chapter 1) about resource-poor model species not being resource poor any longer for any good reason. Then I showed how comparative transcriptomics can drive research of other people: according to ISI Web of Knowledge the ButterflyBase paper has been cited 22 times in the period of January 2008 to August 2010. In the next paper, I identified bottlenecks in NGS and addressed them with the 'Highly Accessed' est2assembly. In another paper, I used new IT and bioinformatic concepts to build GMOD-DBSF, providing the necessary informatic platform for upgrading ButterflyBase. I used all of the above to build ButterflyBase's descendant, InsectaCentral, and attempted to address the curation bottleneck. These resources were used to address specific biological questions. The use of NGS proved to be highly successful, showing that a genome is not required as a reference if a limited amount of wet-lab validation and curation is undertaken. The common denominator of a successful NGS project was a careful project design. Initial pilot studies coupled with bioinformatic consultation would have prevented all of the difficultioes which arose in the case studies. With a good design, access to both a bioinformatic and a wet-lab capability integrative genomics can provide important breakthroughs.

As my co-authors and I wrote in chapter 1: “However for butterflies (or add your taxon here) to fully emerge as ecological and evolutionary genomic models, commitment of the whole research community is required. A concerted effort is crucial to stimulate the development of shared resources and strategies required to turn (your taxon here) into competitive players in the genomics era and to enable a more complete analysis of the questions that have made this group such powerful biological models”. Ignoring for the moment the dreadful length of this sentence, we can recognise that we are looking at the need for a community-wide effort to produce data-type and species-neutral solutions that address the need for bioinformatic support. Subsequently in Chapter 1, we listed resources and requirements which have not been addressed by this thesis: genetic mapping, phenotypes and genetic information, habitat data, evolution and development data. Later we noted that the immediate focus of most research communities is to first generate and curate a

reference sequence, preferably genomic.

As shown in the case studies, the fact that we no longer require a fully ascertained genome, or perhaps any genome at all, is welcoming across the research community despite the increased operational costs. One would keep in mind, however, that with research funding unlikely to meet demand, scientists will still have to select a small number of resource-models in order to understand specific biological phenomena. The major change is, perhaps, that we have fewer limitations in choosing which species these will be and now our criteria can be based on an organisms biology rather than historical precedent. Once a large number of organisms have been investigated for the same biological phenomenon, a new integrative field, such as the one of Systems Biology or Ecological and Evolutionary Functional Genomics will be instrumental in analyzing and synthesizing the results. As we integrate across data types and synthesize between species one possibility remains predictable: our reliance upon computational approaches will increase. As hinted in the case-studies chapter, human curation of a large amount and diverse data will be inevitable but the current research community has yet to develop a system to address this deficiency. It seems that the -omics field is undergoing a colonial period of collecting data from the four corners of the earth. Historical contigency predicts, therefore, the shift away from private collections and the rise of centralized museums and curators. Collections of both sequence and non-sequence data would be most welcome in fields such as population genetics which have until now been forced to limit themselves in theoretical predictions, statistical modelling of simulated datasets or experiments on a limited dataset with very narrow taxonomic sampling. Such stewards of resources have yet, at the time of writing in 2010, to be provided with the necessary financial resources to meet the research community's expectations. Even though a number of funding bodies have begun showing understanding of the need, it is still unclear what is the best approach for provisioning these museums. Agreement, however, exists on the need to have a general and standardized solution driven by the latest developments in IT, computational algorithms and bioinformatic research. Our bioinformatics operations ought to be sufficiently flexible to allow for the realization that methodologies are become rapidly obsolete, we have an increased reliance on technology, and research is increasingly being conducted by large teams or by a community co-ordinated approach. How this paradigm shift affects the research community as a whole remains to be seen. To quote Prof. John Quackenbush (Harvard University): “Genomics has revolutionized biology, but not in the ways that many scientists initially envisioned. While reference genome sequences and catalogues of genes are useful starting points for understanding development and disease, the tools and technology spawned by the genome project have had a far greater impact.”. Indeed, without the

cohesive bioinformatics infrastructure that genome sequencing projects have helped to spawn, no community will be able to upgrade a 'question-model' to a full model species. I hope that this work has laid one more stone towards this goal.

Overall summary – Zusammenfassung

English

Researchers of biology are interested in finding out how biological processes work and how they have come to be, i.e. evolved. We use model systems to infer processes which occur in a larger part of the natural world; these are defined as question-models in this work. Some model systems have traditionally been experimentally tractable organisms with rich resources (resource-models). Before 2006, i.e. prior the start of the work presented herein, genomic resources were scarce for non-model eukaryotic species. In the Ecological and Evolutionary Functional Genomics (EEFG) field we are interested in the so called “emerging models”, i.e. question-models which are using novel technologies to cost-effectively generate the required -omic resources. The main research theme in this work is the building, use and dissemination of transcriptomic data. Next Generation Sequencing (NGS) technologies have made sequence generation more cost-effective but introduced a bioinformatic bottleneck. Therefore, the presented work addresses the bioinformatic bottlenecks relevant in converting a question-model species into a resource-model species using transcriptomics. First, the value and feasibility of this aim is explored and outlined using the emerging model species of *Bicyclus anynana* and *Heliconius* species in paper published in the journal Heredity. Each subsequent chapter addresses one bottleneck: data analysis, data integration and dissemination. A case studies chapters shows the utility of the approach for investigating specific biological phenomena.

The utility of this approach was initially shown in a proof-of-concept paper entitled ButterflyBase published in Nucleic Acids Research. This paper produced a taxon-wide (Lepidoptera: butterflies and moths) online resource with reference transcriptomes prepared from public data found in GenBank. Because the paper was well received (23 citations between January 2008 and October 2010; source: ISI Web of Knowledge, accessed 03 October 2010) subsequent work focused on enhancing the bioinformatic system in order to produce a taxon wide resource for transcriptomics.

The *est2assembly* software, published in the journal BMC Bioinformatics, is a complete platform for producing reference transcriptomes using traditional or Next Generation Sequencing (NGS) technologies. It utilized the standards set out by the General Model Organism Database (GMOD) consortium in order to produce a standardized platform which can be used by small or large laboratories. The end-product is a reference transcriptome with deep-annotations to facilitate data-mining stored in Chado, the relational database format of GMOD. To address the assembly issue, it allows a number of assemblers to be used as plugins and users can choose the assembly meeting their needs using standard indexes such as coverage of a reference sequence dataset.

The GMOD Drupal Bioinformatic Server Framework (GMOD-DBSF), published in the journal Bioinformatics (Oxford), was then built in order to produce a standardized and robust solution to the database and dissemination bottleneck. The paper provided the first three bioinformatic modules for the Drupal Content Management System (CMS). First, gmod_dbsf is an Application Programming Interface module and simplified the programming of bioinformatic Drupal modules. Second, the Drupal Bioinformatic Software Bench (biosoftware_bench) allowed for a rapid and secure deployment of bioinformatic software with the Drupal CMS. An innovative graphical user interface guides both use and administration of the software, including the secure provision of pre-publication datasets. The third module exemplified how our work supports the wider research community by facilitating a review paper by the Lepidoptera community on RNAi experiments.

The *est2assembly* and Drupal modules the basis for preparing a Next Generation online database for an entire taxon: insects. Public data from GenBank and the Short Read Archive were used to generate reference transcriptomes for hundreds of species. These were complemented by secured pre-publication datasets contributed by collaborators. A new bioinformatic software, genes4all, was used to produce 'Centrals': online databases of reference sequence using the Chado database. Innovations such as secured data, the aforementioned biosoftware_bench and graphical visualization of data set a new standard for online genomic resources. Therefore an InsectaCentral was deployed which contains more than one million predicted proteins and is hoped to become a standard resource in the field.

Thanks to NGS, we can more cost-effectively create reference transcriptomes and this work has successfully bridged the bioinformatic gap in relation to transcriptomics. Reference transcriptomes can be used to answer specific biological questions if the appropriate bioinformatic tools for dissemination and analysis are provided. Thus the final chapter deals with the usage of reference transcriptomes for investigating specific biological questions. The end-result of such bioinformatic experiments is usually i) a set of candidate sequences which need to be investigated with traditional hypothesis-driven molecular research; ii) a better understanding of experimental design and iii) a suite of tools which comply with the software-design criteria mentioned above and can be seamlessly utilized by other bioinformaticians. Further, as a result of this thesis, the GMOD consortium is now capable of processing, analyzing and disseminating transcriptomic data without the use of a reference genome (http://gmod.org/*est2assembly*, http://gmod.org/gmod-dbsf, http://gmod.org/InsectaCentral). The work is contained in a total of five research papers, of which four are published in leading journals of the field and one is unpublished. Further, a collection of case studies is included sections of which can appear in future published work.

Deutsch

Eine wichtige Aufgabe der biologischen Forschung ist die Funktionsweies und Entwicklung biologischer Prozesse funktionieren aufzuklären. Zur Beantwortung dieser Fragen werden einfache Modellsysteme (Tierspezies) herangezogen. Daran gewonnene Erkenntnisse werden auf die Prozesse uebertragen die sich in in komplexeren Systemen in der Natur abspielen. Solche Modellsysteme werden als “question-models” bezeichnet da sie der Klärung spezifischer Fragen dienen. Ältere und traditionelle Modellsysteme sind leicht manipulierbare Organismen leicht manipulierbar für die bereits eine Bandbreite an genomischen Resourcen existiert, daher die Bezeichnung “resource-models”. Vor 2006 und somit zu Beginn dieser Arbeit waren genomische Resourcen für eukaryotische Arten die nicht zu den klassischen “resource-models” gehören rar. Heute befasst sich das Feld der Ecological and Evolutionary Functional Genomics (EEFG) mit neuen aufkommenden Modellen: “emerging-models”. Mit Erfindung des Next Generation Sequencing (NGS) wurden Sequenzierungen kosteneffizienter. Dies sind “question-models” die mit Hilfe neuer Technologien bearbeitet werden um möglichst kosteneffektiv die gewünschten genomischen Resourcen bereitstellen und nutzen zu können. Jedoch entstanden mit der neuen Technologie auch bioinformatische Engpässe.

Diese Arbeit untersucht diese bioinformatischen Engpässe die den Schritt vom “question-model” zum “resource-model” erschweren. Der Fokus hier liegt in der Transkriptomdaten. Betreffende Artikel sind in der Fach-zeitschrift “Heredity” erschienen. Einige Kapitel der Dissertation befassen sich mit den bioinformatischen Engpässen in Analyse, Integration und Verteilung von Daten. Ein weiteres Kapitel verdeutlicht in einer Fallstudie die Nützlichkeit der hier vorgestellten Methode bei der Untersuchung spezifischer biologische Phenomene. Darauf aufbauend ergab sich eine ordnungsweite (Lepidoptera: Schmetterlinge & Motten) online-Resource von Referenz-transkriptomen aus aufbereiteten Daten der öffentlichen Datenbank “GenBank”, veröffentlicht unter dem Titel “ButterflyBase” in der Fachzeitschrift “Nucleic Acids Research”. Da der Artikel sehr gut aufgenommen wurde (23 Zitierungen von Januar 2008 bis Oktober 2010. Quelle: ISI WOK, 03. Oktober 2010) konzentrierte sich die weitere Arbeit auf die Entwicklung einer verbesserten bioinformatischen Infrastruktur um Transkriptom-Resourcen für weitere Taxa zu generieren.

Die est2assembly Software, publiziert in der Fachzeitschrift “BMC Bioinformatics”, bedient sich der von der “General Model Organisms Database” (GMOD) vorgegeben Standards und ist die einzige Platform die es ermöglicht Transkriptomprojekte zu standardisieren. Beide Sequenzierungs-techniken werden akzeptiert: die klassische Sanger-Sequenzierung sowie die neue NGS Technologie. Das Endprodukt ist ein Referenztranskriptom versehen mit “deep-annotations” um das Data Mining in Chado (das relationale Datenbankformat von GMOD) zu vereinfachen. Das

Programm erlaubt die Anwendung einer Anzahl von Assemblern als plugins. Die Anwender können eine Sequenzmontage (assembly) generieren die ihren spezifischen Anforderungen entspricht. Dies wird ermöglicht durch die Nutzung von Standardindezes. Um Engpässe in der Datenbank und in der Verteilung der Daten zu überwinden wurde das GMOD-DBSF (GMOD Drupal Bioinformatic Server Framework) geschaffen und in der Fachzeitschrift “Bioinformatics (Oxford)” publiziert. Darin werden 3 bioinformatischen Bausteine für das “Drupal Content Management System” vorgestellt. 1) “gmod_dbsf”, ein “Application Programming Interface”, welches die Programmierung bioinformatischer Drupal-Module vereinfacht. 2) “Drupal Bioinformatic Software Bench”, ermöglicht einen schnellen und geschützten Einsatz der bioinformatischen Software mit Drupal CMS. Ein innovatives “graphical user interface” lenkt Anwendung und Verwaltung der Software, einschließlich der geschützten Bereitstellung unpublizierter Datensätze. 3) mit Hilfe eines experimentellen Moduls wird veranschaulicht wie die vorliegende Arbeit die Forschungsgemeinde bereichert. Als Beispiel wurde ein Review gewählt das sich mit RNAi Experimenten in der Gruppe der Lepidoptera befasst. Die est2assembly- und Drupal-Module bilden die Basis einer Next Generation online Datenbank für eine gesamte Tierklasse: Insekten. Aus öffentlichen Daten konnten Referenztranskriptome für Hunderte von Spezies erstellt werden. Diese werden durch geschützte unpublizierte Datensätze von Kollaborateuren ergänzt. Mit einer neuen Bioinformatiksoftware namens “genes4all” wurde “Centrals” erschaffen: eine online-Datenbank von Referenzsequenzen unter Einbezug der Chado-Datenbasis. Innovationen wie die Sicherung der Daten, die graphische Visualisierung von Datensätzen und “biosoftware_bench” setzen neue Maß-stäbe für genomische Resourcen die online zugängig sind. Im nächsten Schritt wurde “InsectaCentral” erstellt. Diese Datenbank umfasst über eine Million prognostizierte Proteine (“predicted proteins”) und hat das Potential eine Standardresource für Entomologen zu werden.

Die vorliegende Arbeit hat erfolgreich die aus der neuen Sequenziermethode resultierenden bioinformatischen Mängel im Bereich der Transkriptomiks behoben. Referenztrans-kriptome können nun konsultiert werden um biologische Fragen zu beantworten sofern die bioinformatischen Hilfsmittel für Datenweitergabe und Analyse bereit stehen. Damit beschäftigt sich das letzte Katpitel. Das Ergebnis solcher bioinformatischen Experimente besteht gewöhnlich aus i) einem Satz von Kandidatensequenzen welche mit molekularen Techniken auf die Richtigkeit der Eingangshypothese untersucht werden müssen, ii) einem besseren Verständnis des experimentellen Designs und iii) einem Satz von Werkzeugen die sich den oben genannten Software-Design-Kriterien fügen und somit übergangslos von anderen Bioinformatikern genutzt werden können. Als weiteres Resultat dieser Arbeit ist das GMOD Konsortium nun in der Lage Transkriptomdaten zu prozessieren, zu analysieren und zu verteilen ohne ein Referenzgenom einbinden zu müssen.

Bibliography

Ababneh, F., L. S Jermiin, C. Ma, and J. Robinson. 2006. Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics 22(10): 1225.

Abascal, F., R. Zardoya, and D. Posada. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21(9): 2104.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ 1997: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs Nucleic Acids Research, 25:3389-3402.

Arnaiz,O. et al. (2006) ParameciumDB: a community resource that integrates the Paramecium tetraurelia genome sequence with genetic data. Nucleic Acids Research, 35, D439 -D444.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT: Gene Ontology: tool for the unification of biology. Nat Genet 2000, 25:25-29.

Bae, J. S., I. Kim, H. D. Sohn, and B. R. Jin. 2004. The mitochondrial genome of the firefly, Pyrocoelia rufa: complete DNA sequence, genome organization, and phylogenetic analysis with other insects. Molecular phylogenetics and evolution 32, no. 3: 978-985.

Baird, Nathan A., Paul D. Etter, Tressa S. Atwood, Mark C. Currey, Anthony L. Shiver, Zachary A. Lewis, Eric U. Selker, William A. Cresko, and Eric A. Johnson. 2008. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. PLoS ONE 3, no. 10. doi:10.1371/journal.pone.0003376.

Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M: The universal protein resource (UniProt). Nucleic Acids Res 2005, 33:D154-159.

Bairoch A: The ENZYME database in 2000 Nucleic Acids Res 2000, 28:304-305

Baxter, S. et al., BA, 2010. Genomic hotspots for adaptation: the population genetics of Müllerian mimicry in the Heliconius melpomene clade. PLoS Genetics, 6(2), p.e1000794.

Beldade, P., Brakefield, P.M. & Long, A.D., 2005. Generating phenotypic variation: prospects from" evo-devo" research on Bicyclus anynana wing patterns. Evolution & Development, 7(2), pp.101-107.

Beldade P, McMillan WO, Papanicolaou A: Butterfly genomics eclosing Heredity 2008, 100:150-157.

Beldade,P., Rudd,S., Gruber,J.D. and Long,A.D. (2006) A wing expressed sequence tag resource for Bicyclus anynana butterflies, an evo-devo model. BMC Genomics, 7, 130.

Beldade P, Saenko SV, Pul N, Long AD: A Gene-Based Linkage Map for Bicyclus anynana Butterflies Allows for a Comprehensive Analysis of Synteny with the Lepidopteran Reference Genome PLoS Genet 2009, 5:e1000366.

Benjamini, Y. & Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological), 57(1), pp.289-300.

Bextine B, Tuan S, Shaikh H, Blua M, Miller TA: Evaluation of Methods for Extracting Xylella fastidiosa DNA from the Glassy-Winged Sharpshooter J Econ Entomol 2004, 97:757-763.

Bickel, P.J. et al., An Overview of Recent Developments in Genomics and the Statistical Methods that Bear on Them. Available at: http://www.stat.berkeley.edu/~bickel/Bickel et al 2009.pdf.

Boguski MS, Lowe TMJ, Tolstoshev CM: dbEST— database for "expressed sequence tags" Nat Genet 1993, 4:332-333.

Bouck, A. & Vision, T., 2007. The molecular ecologist's guide to expressed sequence tags. Molecular Ecology, 16(5), pp.907-24.

Bowker, A. H. 1948. A test for symmetry in contingency tables. Journal of the american statistical association 43, no. 244: 572-574.

Brakefield PM, Gates J, Keys D, Kesbeke F, Wijngaarden PJ, Monteiro A, French V, Carroll SB: Development, plasticity and evolution of butterfly eyespot patterns Nature 1996, 384: 236-242.

Bretman, A., D. Newcombe, and T Tregenza. 2009. Promiscuous females avoid inbreeding by controlling sperm storage. Molecular Ecology 18, no. 16: 3340–3345.

Bretman, A., N. Wedell, and T. Tregenza. 2004. Molecular evidence of post-copulatory inbreeding avoidance in the field cricket Gryllus bimaculatus. Proceedings of the Royal Society B: Biological Sciences 271, no. 1535 (1): 159-164. doi:10.1098/rspb.2003.2563.

Bretman, A., R. Rodríguez-Muñoz, and T. Tregenza. 2006. Male dominance determines female egg laying rate in crickets. Biology Letters 2, no. 3 (9): 409-411. doi:10.1098/rsbl.2006.0493.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular Biology and Evolution 17, no. 4: 540.

Cheng,T.C., Xia,Q.Y., Qian,J.F., Liu,C., Lin,Y., Zha,X.F. and Xiang,Z.H. (2004) Mining single

nucleotide polymorphisms from EST data of silkworm, Bombyx mori, inbred strain Dazao. Insect Biochem. Mol. Biol., 34, 523–530.

Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WEG, Wetter T, Suhai S: Using the miraEST Assembler for Reliable and Automated mRNA Transcript Assembly and SNP Detection in Sequenced ESTs Genome Res 2004, 14:1147-1159.

Chevreux B, Wetter T, Suhai S: Genome sequence assembly using trace signals and additional sequence information In Proc German Conf Bioinformatics 1999, 99:45–56.

Clark, A.G., M.B. Eisen, D.R. Smith, C.M. Bergman, B. Oliver, T.A. Markow, T.C. Kaufman, M. Kellis, W. Gelbart, and V.N. Iyer. 2007. Evolution of genes and genomes on the Drosophila phylogeny. Nature 450, no. 7167: 203-218.

Clarke, C. A., and P. M. Sheppard. 1963. Interactions between major genes and polygenes in the determination of the mimetic patterns of Papilio dardanus. Evolution: 404-413.

Clark, R., S.M. Brown, S.C. Collins, C.D. Jiggins, D.G. Heckel, and A.P. Vogler. 2008. Colour pattern specification in the Mocker swallowtail Papilio dardanus: the transcription factor invected is a candidate for the mimicry locus H. Proceedings of the Royal Society B: Biological Sciences 275, no. 1639 (5): 1181-1188. doi:10.1098/rspb.2007.1762.

Collins, F.S., Morgan, M. & Patrinos, A., 2003. The Human Genome Project: lessons from large-scale biology. Science, 300(5617), pp.286-290.

Crawford, D., 2001. Functional genomics does not have to be limited to a few select organisms. Genome Biology, 2(1).

Crick, F., 1970. Central dogma of molecular biology. Nature, 227(5258), pp.561-563.

Crosby,M.A., Goodman,J.L., Strelets,V.B., Zhang,P. and Gelbart,W.M. (2007) FlyBase: genomes by the dozen. Nucleic Acids Res., 35, D486–D491.

Day,A. et al. (2007) Celsius: a community resource for Affymetrix microarray data. Genome Biology, 8, R112.

Delseny, M., Han, B. & Hsing, Y.I., High throughput DNA sequencing: The new sequencing revolution. Plant Science, In Press,.

DeRisi, J., L. Penland, P. O. Brown, M. L. Bittner, P. S. Meltzer, M. Ray, Y. Chen, Y. A. Su, and J. M. Trent. 1996. Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nature genetics 14, no. 4: 457-460.

Drysdale,R.A. and Crosby,M.A. (2005) FlyBase: genes and gene models. Nucleic Acids Research, 33, D390-395.

Duan, J., R. Li, D. Cheng, W. Fan, X. Zha, T. Cheng, Y. Wu, J. Wang, K. Mita, and Z. Xiang. 2010. SilkDB v2. 0: a platform for silkworm (Bombyx mori) genome biology. Nucleic Acids Research 38: D453.

Durbin, R. et al., 2002. Biological sequence analysis, Cambridge university press Cambridge, UK:

Dutilh, B. E., V. van Noort, R. T. J. van der Heijden, T. Boekhout, B. Snel, and M. A. Huynen. 2007. Assessment of phylogenomic and orthology approaches for phylogenetic inference. Bioinformatics 23, no. 7: 815.

Eddy, S. R. 2000. HMMER: Profile hidden Markov models for biological sequence analysis. Washington University School of Medicine, St Louis, MO (http://hmmer. wustl. edu/).

Efron, B., and R. Tibshirani. 2002. Empirical Bayes methods and false discovery rates for microarrays. Genetic Epidemiology 23, no. 1: 70-86.

Eilbeck, K. et al., 2005. The Sequence Ontology: a tool for the unification of genome annotations. Genome Biology, 6(5), p.R44.

Emmersen J, Rudd S, Mewes HW, Tetko IV: Separation of sequences from host–pathogen interface using triplet nucleotide frequencies Fungal Genet Biol 2007, 44:231-241.

Enright, A.J., Van Dongen, S. & Ouzounis, C.A., 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research, 30(7), p.1575.

Ewing B, Green P: Analysis of expressed sequence tags indicates 35,000 human genes Nat Genet 2000, 25:232-234.

Ewing,B., Hillier,L., Wendl,M. and Green,P. (1998) Basecalling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res., 8, 175–185.

Eyre-Walker, A. 1996. Synonymous codon bias is related to gene length in Escherichia coli: selection for translational accuracy? Molecular biology and evolution 13, no. 6: 864.

Ferguson L, Lee SF, Chamberlain N, Nadea N, Joron M, Baxter S, Wilkinson P, Papanicolaou A, Kumar S, Thuan-Jin Clark R, Davidson C, Glithero R, Beasle H, Vogel H, Ffrench-Constant R H, Jiggins CD: Characterization of a hotspot for mimicry: assembly of a butterfly wing transcriptome to genomic sequence at the hmyb/sb locus. Mol Ecol in press

Feyereisen, R. 2006. Evolution of insect P450. Biochemical Society Transactions 34: 1252-1255.

Filipski, J. 1987. Correlation between molecular clock ticking, codon usage, fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. FEBS letters 217, no. 2: 184-186.

Foster, P. G., and D. A. Hickey. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. Journal of Molecular Evolution 48, no. 3: 284-290.

Friedel CC, Jahn KHV, Sommer S, Rudd S, Mewes HW, Tetko IV: Support vector machines for separation of mixed plant-pathogen EST collections based on codon usage Bioinformatics 2005, 21:1383-1388.

Fukunishi,Y. and Hayashizaki,Y. (2001) Amino acid translation program for full-length cDNA sequences with frameshift errors. Physiol. Genomics, 5, 81–87.

Futahashi, R., and H. Fujiwara. 2005. Melanin-synthesis enzymes coregulate stage-specific larval cuticular markings in the swallowtail butterfly, Papilio xuthus. Development genes and evolution 215, no. 10: 519-529.

Futahashi, R., and H. Fujiwara. 2006. Expression of one isoform of GTP cyclohydrolase I coincides with the larval black markings of the swallowtail butterfly, Papilio xuthus. Insect biochemistry and molecular biology 36, no. 1: 63-70.

Futahashi, R., and H. Fujiwara. 2007. Regulation of 20-hydroxyecdysone on the larval pigmentation and the expression of melanin synthesis enzymes and yellow gene of the swallowtail butterfly, Papilio xuthus. Insect biochemistry and molecular biology 37, no. 8: 855-864.

Futahashi, R., J. Sato, Y. Meng, S. Okamoto, T. Daimon, K. Yamamoto, Y. Suetsugu, J. Narukawa, H. Takahashi, and Y. Banno. 2008. yellow and ebony Are the Responsible Genes for the Larval Color Mutants of the Silkworm Bombyx mori. Genetics 180, no. 4: 1995.

Futahashi, R., Y. Banno, and H. Fujiwara. 2010. Caterpillar color patterns are determined by a two-phase melanin gene prepatterning process: new evidence from tan and laccase2. Evolution & Development 12, no. 2: 157-167.

Gauthier, J.P. et al., 2007. AphidBase: a database for aphid genomic resources. Bioinformatics, 23(6), p.783.

Gene Ontology Consortium. (2006) The gene ontology (GO) project in 2006. Nucleic Acids Res., 34, D322–D326.

Giardine,B. (2005) Galaxy: A platform for interactive large-scale genome analysis. Genome Research, 15, 1451-1455.

Goldsmith MR, Shimada T, Abe H: The genetics and genomics of the silkworm, Bombyx mori Annu Rev Entomol 2005, 50:71-100.

Gray, D. A. 2005. Does courtship behavior contribute to species-level reproductive isolation in field crickets? Behavioral Ecology 16, no. 1: 201.

Greenbaum, D. et al., 2001. Interrelating different types of genomic data, from proteome to secretome:'oming in on function. Genome Research, 11(9), p.1463.

Green, P., 1996. phrap. Genome Center, University of Washington.

Guindon, S., F. Delsuc, J. F. Dufayard, and O. Gascuel. 2009. Estimating Maximum Likelihood Phylogenies with PhyML. Methods in molecular biology (Clifton, NJ) 537: 113.

Hahn, M.W., Han, M.V. & Han, S.-G., 2007. Gene Family Evolution across 12 Drosophila Genomes. PLoS Genetics, 3(11), p.e197.

Harismendy O, Frazer K: Method for improving sequence coverage uniformity of targeted genomic intervals amplified by LR-PCR using Illumina GA sequencing-by-synthesis technology BioTechniques 2009, 46:229.

Heckel DG, Gahan LJ, Daly JC, Trowell S: A genomic approach to understanding Heliothis and Helicoverpa resistance to chemical and biological insecticides Philos Trans R Soc Lond, B, Biol Sci 1998, 353:1713-1722.

Heinicke,S. et al. (2007) The Princeton Protein Orthology Database (P-POD): a comparative genomics analysis tool for biologists. PLoS One, 2.

Honegger, HW. 1981. Three different diel rhythms of the calling song in the cricket, Gryllus campestris, and their control mechanisms*. Physiological Entomology 6, no. 3: 289-296.

Huang X, Madan A: CAP3: A DNA sequence assembly program Genome Res 1999, 9:868-877.

Hubbard,T. et al. (2002) The Ensembl genome database project. Nucleic Acids Research, 30, 38.

Huntley,D., Baldo,A., Johri,S. and Sergot,M. (2006) SEAN: SNP prediction and display program utilizing EST sequence clusters. Bioinformatics, 22, 495–496.

Iseli,C., Jongeneel,C.V. and Bucher,P. (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc. Int. Conf. Intell. Syst. Mol. Biol., 138–148, http://www.ch.embnet.org/software/ESTScan.html.

Jang, Yikweon, and H. Carl Gerhardt. 2006. Divergence in female calling song discrimination between sympatric and allopatric populations of the southern wood cricket Gryllus fultoni

(Orthoptera: Gryllidae). Behavioral Ecology and Sociobiology 60, no. 2 (2): 150-158. doi:10.1007/s00265-005-0151-3.

Jermiin, L.S., V. Jayaswal, F. Ababneh, and J. Robinson. 2008. Phylogenetic model evaluation. In Bioinformatics, ed. Jonathan M. Keith. Vol. 452. Totowa, NJ: Humana Press. http://www.springerlink.com/index/10.1007/978-1-60327-159-2.

Ji G, Zheng J, Shen Y, Wu X, Jiang R, Lin Y, Loke J, Davis K, Reese G, Li Q: Predictive modeling of plant messenger RNA polyadenylation sites BMC Bioinformatics 2007, 8:43

Johnson, N. F., and C. A. Triplehorn. 2004. Borror and DeLong's Introduction to the Study of Insects. Brooks/Cole Publishing Company.

Joron, M. et al., 2006. Heliconius wing patterns: an evo-devo model for understanding phenotypic diversity. Heredity, 97(3), pp.157-167.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: Repbase Update, a database of eukaryotic repetitive elements Cytogenet Genome Res 2005, 110: 462-467.

Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res., 34, D354–D357.

Kanehisa M, Goto S: KEGG: Kyoto Encyclopedia of Genes and Genomes Nucleic Acids Res 2000, 28:27-30.

Kent,W. et al. (2002) The human genome browser at UCSC. Genome Research, 12, 996 - 1006.

Kitano, H., 2002. Systems biology: a brief overview. Science, 295(5560), p.1662.

Kjer, K. M. 2004. Aligned 18S and insect phylogeny. Systematic Biology 53, no. 3: 506.

Landais, I., M. Ogliastro, K. Mita, J. Nohata, M. Lopez-Ferber, M. Duonor-Cerutti, T. Shimada, P. Fournier, and G. Devauchelle. 2003. Annotation pattern of ESTs from Spodoptera frugiperda Sf9 cells and analysis of the ribosomal protein genes reveal insect-specific features and unexpectedly low codon usage bias. Bioinformatics 19, no. 18: 2343.

Lander, E.S. et al., 2001. Initial sequencing and analysis of the human genome. Nature, 409(6822), pp.860-921.

Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10, no. 3: R25.

Langmead, B., M. Schatz, J. Lin, M. Pop, and S. Salzberg. 2009. Searching for SNPs with cloud

computing. Genome Biology 10, no. 11: R134.

Lawson D, Arensburger P, Atkinson P, Besansky NJ, Bruggner RV, Butler R, Campbell KS, Christophides GK, Christley S, Dialynas E: VectorBase: a data resource for invertebrate vector genomics. Nucleic Acids Res 2009, 37:D583-587.

Lawson D, Arensburger P, Atkinson P, Besansky NJ, Bruggner RV, Butler R, Campbell KS, Christophides GK, Christley S, Dialynas E: VectorBase: a home for invertebrate vectors of human pathogens. Nucleic Acids Res 2007, 35:D503-505.

Lazebnik, Y., 2004. Can a biologist fix a radio?—Or, what I learned while studying apoptosis. Biochemistry (Moscow), 69(12), pp.1403-1406.

Li W, Godzik A: Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences Bioinformatics 2006, 22:1658-9.

Longhorn, S.J., Pohl, H.W. & Vogler, A.P., 2010. Ribosomal protein genes of holometabolan insects reject the Halteria, instead revealing a close affinity of Strepsiptera with Coleoptera. Molecular Phylogenetics and Evolution.

Lyne,R. et al. (2007) FlyMine: an integrated database for Drosophila and Anopheles genomics. Genome Biology, 8, R129.

Mangone, M. et al., 2010. The Landscape of C. elegans 3'UTRs. Science.

Marioni, J. C., C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome research 18, no. 9: 1509.

McComish, B. J., S. F. K. Hills, P. J. Biggs, and D. Penny. 2010. Index-Free De Novo Assembly and Deconvolution of Mixed Mitochondrial Genomes. Genome Biology and Evolution 2, no. 0 (5): 410-424. doi:10.1093/gbe/evq029.

Miao,X.X., Xub,S.J., Li,M.H., Li,M.W., Huang,J.H., Dai,F.Y., Marino,S.W., Mills,D.R., Zeng,P. et al. (2005) Simple sequence repeat-based consensus linkage map of Bombyx mori. Proc. Natl Acad. Sci. USA, 102, 16303–16308.

Miller, Jason R et al., 2008. Aggressive assembly of pyrosequencing reads with mates. Bioinformatics, 24(24), pp.2818-24.

Miller, J. R, Koren, S. & Sutton, G., 2010. Assembly algorithms for next-generation sequencing data. Genomics.

Misof, B., O. Niehuis, I. Bischoff, A. Rickert, D. Erpenbeck, and A. Staniczek. 2007. Towards an 18S phylogeny of hexapods: accounting for group-specific character covariance in optimized mixed nucleotide/doublet models. Zoology 110, no. 5: 409-429.

Mita,K., Kasahara,M., Sasaki,S., Nagayasu,Y., Yamada,T., Kanamori,H., Namiki,N., Kitagawa,M., Yamashita,H. et al. (2004) The genome sequence of silkworm, Bombyx mori. DNA Res., 11, 27–35.

Mita,K., Morimyo,M., Okano,K., Koike,Y., Nohata,J., Kawasaki,H., Kadono-Okuda,K., Yamamoto,K., Suzuki,M.G. et al. (2003) The construction of an EST database for Bombyx mori and its application. Proc. Natl Acad. Sci. USA, 100, 14121–14126.

Mitchell-Olds, T., Feder, M. & Wray, G., Evolutionary and ecological functional genomics. Heredity, 100(2), pp.101-102.

Monteiro, A. & Prudic, K.M., 2010. Multiple approaches to study color pattern evolution in butterflies. Trends in Evolutionary Biology, 2(1), p.e2.

Mungall,C.J. and Emmert,D.B. FlyBase Consortium (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. Bioinformatics, 23, i337–i346.

Mungall, C.J. & Emmert, D.B., 2007. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. Bioinformatics, 23(13), p.i337.

Murray, A.W., 2000. Whither genomics? Genome Biology, 1(1).

Negre,V., Hotelier,T., Volkoff,A.N., Gimenez,S., Cousserans,F., Mita,K., Sabau,X., Rocher,J., Lopez-Ferber,M. et al. (2006) SPODOBASE: an EST database for the lepidopteran crop pest Spodoptera. BMC Bioinformatics, 7, 322.

Ning, Z., Cox, A.J. & Mullikin, J.C., 2001. SSAHA: A fast search method for large DNA databases. Genome Research, 11(10), p.1725.

Oakeshott, J. G., R. M. Johnson, M. R. Berenbaum, H. Ranson, A. S. Cristino, and C. Claudianos. 2010. Metabolic enzymes associated with xenobiotic and chemosensory responses in Nasonia vitripennis. Insect Molecular Biology 19, no. 1: 147-163.

O'Brien, K.P., Remm, M. & Sonnhammer, E.L.L., 2005. Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Research, 33, p.D476.

O'Connor,B. et al. (2008) GMODWeb: a web framework for the generic model organism database.

Genome Biology, 9, R102.

Oinn,T. et al. (2004) Taverna: a tool for the composition and enactment of bioinformatics work flows. Bioinformatics 20, 3045-3054.

O'Neil, S. T., J. D. K. Dzurisin, R. D. Carmichael, N. F. Lobo, S. J. Emrich, and J. J. Hellmann. 2010. Population-level transcriptome sequencing of nonmodel organisms Erynnis propertius and Papilio zelicaon. BMC genomics 11, no. 1: 310.

Ozsolak, F., A. R. Platt, D. R. Jones, J. G. Reifenberger, L. E. Sass, P. McInerney, J. F. Thompson, J. Bowers, M. Jarosz, and P. M. Milos. 2009. Direct RNA sequencing. Nature 461, no. 7265: 814-818.

Papanicolaou A, Gebauer-Jung S, Blaxter ML, McMillan WO, Jiggins, CD: ButterflyBase: a platform for lepidopteran genomics Nucleic Acids Res 2008, 36:D582-587.

Papanicolaou,A., Joron,M., McMillan,W.O., Blaxter,M.L. and Jiggins,C.D. (2005) Genomic tools and cDNA derived markers for butterflies. Mol Ecol., 14, 2883–2897.

Papanicolaou, Alexie et al., 2009. Next generation transcriptomes for next generation genomes using est2assembly. BMC Bioinformatics, 10(1), p.447.

Paquola AC, Nishyiama Jr MY, Reis EM, da Silva AM, Verjovski-Almeida S: ESTWeb: bioinformatics services for EST sequencing projects Bioinformatics 2003, 19:1587-1587.

Parkinson,J., Anthony,A., Wasmuth,J., Schmid,R., Hedley,A. and Blaxter,M. (2004) PartiGene—constructing partial genomes. Bioinformatics, 20, 1398–1404.

Parkinson,J., Guiliano,D.B. and Blaxter,M. (2002) Making sense of EST sequences by CLOBBing them. BMC Bioinformatics, 3, 31.

Pauchet, Y., P. Wilkinson, H. Vogel, D. R. Nelson, S. E. Reynolds, D. G. Heckel, and others. 2009. Pyrosequencing the Manduca sexta larval midgut transcriptome: messages for digestion, detoxification and defence. Insect Molecular Biology 19, no. 1: 61–75.

Pauchet Y, Wilkinson P, van Munster M, Augustin S, Pauron D, Ffrench-Constant RH: Pyrosequencing of the midgut transcriptome of the poplar leaf beetle Chrysomela tremulae reveals new gene families in Coleoptera. Insect Biochem Mol Biol. 2009, 39:403-13

Pearson WR, Wood T, Zhang Z, Miller W: Comparison of DNA sequences with protein sequences Genomics 1997, 46:24-36.

Philippe, H., and M. J. Telford. 2006. Large-scale sequencing and the new animal phylogeny.

Trends in Ecology & Evolution 21, no. 11: 614-620.

Phred, Phrap, and Consed [http://www.phrap.com]

Piel, W. H., M. J. Donoghue, and M. J. Sanderson. 2000. TreeBASE: A database of phylogenetic information. In Proceedings of the 2nd International Workshop of Species 2000.

Pringle,E.G., Baxter,S.W., Webster,C.L., Papanicolaou,A., Lee,S.F. and Jiggins,C.D. (2007) Synteny and chromosome evolution in the Lepidoptera: evidence from mapping in Heliconius melpomene. Genetics, 177, 417–426.

Pruitt KD, Tatusova T, Maglott DR: NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins Nucleic Acids Res 2007, 35:D61-65.

Rajaram, S., and Y. Oono. 2010. NeatMap- non-clustering heat map alternatives in R. BMC bioinformatics 11, no. 1: 45.

Rausher MD: Natural selection and the evolution of plant insect interactions In Insect chemical ecology: an evolutionary approach. Edited by Rausher MD, Isman MB. New York: Chapman & Hall; 1992:20-88.

Reed, R. D., P. H. Chen, and H. F. Nijhout. 2007. Cryptic variation in butterfly eyespot development: the importance of sample size in gene expression studies. Evolution & Development 9, no. 1: 2-9.

Regier, J. C., A. Zwick, M. P. Cummings, A. Y. Kawahara, S. Cho, S. Weller, A. Roe, J. Baixeras, J. W. Brown, and C. Parr. 2009. Toward reconstructing the evolution of advanced moths and butterflies(Lepidoptera: Ditrysia): an initial molecular study. BMC Evolutionary Biology 9, no. 1: 280.

RepeatMasker [http://www.repeatmasker]

Rice, P., I. Longden, and A. Bleasby. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16: 276 - 277.

Robinson, M. D, and G. K Smyth. 2007. Moderated statistical tests for assessing differences in tag abundance. Bioinformatics 23, no. 21: 2881.

Robinson, M.D. & Oshlack, A., 2010. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology, 11(3), p.R25.

Rocha, E. P. C., and A. Danchin. 2002. Base composition bias might result from competition for metabolic resources. TRENDS in Genetics 18, no. 6: 291-294.

Ronaghi, M., 2001. Pyrosequencing sheds light on DNA sequencing. Genome Research, 11(1), p.3.

Rothberg, J.M. & Leamon, J.H., 2008. The development and impact of 454 sequencing. Nature Biotechnology, 26(10), pp.1117-1124.

Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol. Biol., 132,365–386.

Rudd S: Expressed sequence tags: alternative or complement to whole genome sequences? Trends Plant Sci 2003, 8:321-329.

Ruzanov, P., and D. L Riddle. 2010. Deep SAGE analysis of the Caenorhabditis elegans transcriptome. Nucleic Acids Research 38, no. 10: 3252.

Salzberg,S.L. (2007) Genome re-annotation: a wiki solution? Genome Biol., 8, 102.

Sambrook, J., E. F. Fritsch, and T. Maniatis. 1989. Molecular cloning. Cold Spring Harbor Laboratory Press Cold Spring Harbor, NY.

Samways, M. J. 1993. Insects in biodiversity conservation: some perspectives and directives. Biodiversity and conservation 2, no. 3: 258-282.

Sato, K., T. M. Matsunaga, R. Futahashi, T. Kojima, K. Mita, Y. Banno, and H. Fujiwara. 2008. Positional cloning of a Bombyx wingless locus flugellos (fl) reveals a crucial role for fringe that is specific for wing morphogenesis. Genetics 179, no. 2: 875.

Sauer, U., Heinemann, M. & Zamboni, N., 2007. GENETICS: Getting Closer to the Whole Picture. Science, 316(5824), pp.550-551.

Savard, J., D. Tautz, S. Richards, G. M. Weinstock, R. A. Gibbs, J. H. Werren, H. Tettelin, and M. J. Lercher. 2006. Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of Holometabolous insects. Genome Research 16, no. 11 (10): 1334-1338. doi:10.1101/gr.5204306.

Schmid,R. and Blaxter,M.L. (2008) annot8r: GO, EC and KEGG annotation of EST datasets. BMC Bioinformatics, 9, 180.

Schuster SC: Next-generation sequencing transforms today's biology Nat Methods 2008, 5:16-18.

Scriber, J. M., M. H. Evans, and D. B. Ritland. 1986. Hybridization as a causal mechanism of mixed color broods and unusual color morphs of female offspring in the eastern tiger swallowtail butterflies, Papilio glaucus. Evolutionary genetics of invertebrate behavior: progress and prospects: 119.

Scriber, J. M., R. H. Hagen, and R. C. Lederhouse. 1996. Genetics of mimicry in the tiger swallowtail butterflies, Papilio glaucus and P. canadensis (Lepidoptera: Papilionidae). Evolution 50, no. 1: 222-236.

SFF extract [http://bioinf.comav.upv.es/sff_extract/]

Sharp, P. M., and G. Matassi. 1994. Codon usage and genome evolution. Current Opinion in Genetics & Development 4, no. 6: 851-860.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: dbSNP: the NCBI database of genetic variation Nucleic Acids Res 2001, 29:308-311.

Shevchenko,A., Sunyaev,S., Loboda,A., Shevchenko,A., Bork,P., Ens,W. and Standing,K.G. (2001) Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. Anal. Chem., 73, 1917–1926.

Shimomura, M., Y. Shimizu, A. B. A. Sasanuma Si, Y. Nagamura, K. Mita, and T. Sasaki. 2004. KAIKOGAAS: An automated annotation System for silkworm genome. In Genome Inform. Vol. 181.

Shirataki, H., R. Futahashi, and H. Fujiwara. 2010. Species-specific coordinated gene expression and trans-regulation of larval color pattern in three swallowtail butterflies. Evolution & Development 12, no. 3: 305-314.

Simmons, L. W. 2004. Genotypic variation in calling song and female preferences of the field cricket Teleogryllus oceanicus. Animal behaviour 68, no. 2: 313-322.

Sjolander, K., 2004. Phylogenomic inference of protein molecular function: advances and challenges. Bioinformatics, 20(2), p.170.

Slonim, D.K., 2002. From patterns to pathways: gene expression data analysis comes of age. Nature Genetics, 32, pp.502-508.

Smedley,D. et al. (2009) BioMart–biological queries made easy. BMC Genomics, 10, 22.

Smith, V., S. Rycroft, K. Harman, B. Scott, and D. Roberts. 2009. Scratchpads: a data-publishing framework to build, share and manage information on the diversity of life. BMC Bioinformatics 10, no. 14: S6.

Snyder, M. J, J. K Walding, and R. Feyereisen. 1995. Glutathione S-transferases from larval Manduca sexta midgut: sequence of two cDNAs and enzyme induction. Insect biochemistry and molecular biology 25, no. 4: 455–465.

Snyder, M. J., J. L. Stevens, J. F. Andersen, and R. Feyereisen. 1995. Expression of cytochrome P450 genes of the CYP4 family in midgut and fat body of the tobacco hornworm, Manduca sexta. Archives of Biochemistry and Biophysics 321, no. 1: 13-20.

Solignac M, Zhang L, Mougel F, Li B, Vautrin D, Monnerot M, Cornuet JM, Worley KC, Weinstock GM, Gibbs RA: The genome of Apis mellifera: dialog between linkage mapping and sequence assembly. Genome Biol 2007, 8:403.

Sommer, D.D. et al., 2007. Minimus: a fast, lightweight genome assembler. BMC Bioinformatics, 8(1), p.64.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H: The Bioperl toolkit: Perl modules for the life sciences. Genome Res 2002, 12:1611-1618.

Stalker, J. et al., 2004. The Ensembl Web site: mechanics of a genome browser. Genome Research, 14(5), p.951.

Stamatakis, A. & Alachiotis, N., 2010. Time and memory efficient likelihood-based tree searches on phylogenomic alignments with missing data. Bioinformatics, 26(12), p.i132.

Stamatakis,A. et al. (2008) A rapid bootstrap algorithm for the RAxML Web servers. Systematic Biology, 57, 758-771.

Stamatakis, A., M. Ott, and T. Ludwig. 2005. Raxml-omp: An efficient program for phylogenetic inference on smps. Parallel Computing Technologies: 288-302.

Stark, A., M. F. Lin, P. Kheradpour, J. S. Pedersen, L. Parts, J. W. Carlson, M. A. Crosby, M. D. Rasmussen, S. Roy, and A. N. Deoras. 2007. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature 450, no. 7167: 219-232.

Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S: The Generic Genome Browser: A Building Block for a Model Organism System Database. Genome Res 2002, 12:1599-1610.

Stevens, J. L., M. J. Snyder, J. F. Koener, and R. Feyereisen. 2000. Inducible P450s of the CYP9 family from larval Manduca sexta midgut. Insect Biochemistry and Molecular Biology 30, no. 7: 559-568.

Stoeckert, C. J., H. C. Causton, and C. A. Ball. 2002. Microarray databases: standards and ontologies. Nature Genetics 32: 469-473.

Storey, J. D, and R. Tibshirani. 2003. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America 100, no. 16: 9440.

Strimmer, K. 2008a. fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. Bioinformatics 24, no. 12: 1461.

Strimmer, K. 2008b. A unified approach to false discovery rate estimation. BMC bioinformatics 9, no. 1: 303.

Stuart, A. 1955. A test for homogeneity of the marginal distributions in a two-way classification. Biometrika 42, no. 3: 412.

Team, R.D.C., 2009. R: A Language and Environment for Statistical Computing, Vienna, Austria. Available at: http://www.r-project.org.

Terenius, O. et al., 2010. RNA interference in Lepidoptera: an overview of successful and unsuccessful studies and implications for experimental design. Journal of Insect Physiology, (doi:10.1016/j.jinsphys.2010.11.006).

Thain, D. et al. (2005) Distributed Computing in Practice: The Condor Experience. Concurrency and Computation, 17, 323-356.

Thalamuthu, A., I. Mukhopadhyay, X. Zheng, and G. C. Tseng. 2006. Evaluation and comparison of gene clustering methods in microarray analysis. Bioinformatics 22, no. 19: 2405.

The Nasonia Genome Working Group, J. H. Werren, S. Richards, C. A. Desjardins, O. Niehuis, J. Gadau, J. K. Colbourne, et al. 2010. Functional and Evolutionary Insights from the Genomes of Three Parasitoid Nasonia Species. Science 327, no. 5963 (1): 343-348. doi:10.1126/science.1178028.

Thiel,T., Michalek,W., Varshney,R. and Graner,A. (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). Theor. Appl. Genet., 106, 411–422.

Thomson RC, Shedlock AM, Edwards SV, Shaffer HB: Developing markers for multilocus phylogenetics in non-model organisms: A test case with turtles Mol Phylogenet Evol 2008, 49:514-525.

Van Straalen NM, Roelofs D: An introduction to ecological genomics Oxford University Press; 2006.

van't Hof, A. E., and I. J. Saccheri. 2010. Industrial Melanism in the Peppered Moth Is Not

Associated with Genetic Variation in Canonical Melanisation Gene Candidates. PloS one 5, no. 5: e10889.

Venter, J.C. et al., 2001. The sequence of the human genome. Science, 291(5507), p.1304.

Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, Marden JH: Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing Mol Ecol 2008, 17:1636-47.

Vieira, F., A. Sánchez-Gracia, and J. Rozas. 2007. Comparative genomic analysis of the odorant-binding protein family in 12 Drosophila genomes: purifying selection and birth-and-death evolution. Genome Biology 8, no. 11: R235.

Wang J, Xia Q, He X, Dai M, Ruan J, Chen J, Yu G, Yuan H, Hu Y, Li R: SilkDB: a knowledgebase for silkworm biology and genomics. Nucleic Acids Res 2005, 33:D399.

Wang L, Wang S, Li Y, Paradesi MSR, Brown SJ: BeetleBase: the model organism database for Tribolium castaneum Nucleic Acids Res 2007, 35:D476-479.

Wang, Y. & Gu, X., 2001. Functional divergence in the caspase gene family and altered functional constraints: statistical analysis and prediction. Genetics, 158(3), p.1311.

Wang, Z., M. Gerstein, and M. Snyder. 2008. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics.

Wasmuth,J.D. and Blaxter,M.L. (2004) prot4EST: translating expressed sequence tags from neglected genomes. BMC Bioinformatics, 5, 187.

Wheeler, W. C., P. Cartwright, and C. Y. Hayashi. 1993. Arthropod phylogeny: a combined approach. Cladistics 9, no. 1: 1-39.

Whiting, M. F. 2002. Phylogeny of the holometabolous insect orders: molecular evidence. Zoologica Scripta 31, no. 1: 3-15.

Whiting, M. F., J. C. Carpenter, Q. D. Wheeler, and W. C. Wheeler. 1997. The Strepsiptera problem: phylogeny of the holometabolous insect orders inferred from 18S and 28S ribosomal DNA sequences and morphology. Systematic Biology 46, no. 1: 1.

Wiegmann, B.M. et al., 2009. Single-copy nuclear genes resolve the phylogeny of the holometabolous insects. BMC Biology, 7(1), p.34.

Wilson, R.J., Goodman, J.L. & Strelets, V.B., 2007. FlyBase: integration and improvements to query tools. Nucleic Acids Research.

Wittkopp, P. J, B. L Williams, J. E Selegue, and S. B Carroll. 2003. Drosophila pigmentation evolution: Divergent genotypes underlying convergent phenotypes. Proceedings of the National Academy of Sciences of the United States of America 100, no. 4 (February): 1808-1813.

Wolf, Y. I., I. B. Rogozin, and E. V. Koonin. 2004. Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. Genome Research 14, no. 1: 29.

Woodhead,M., Russell,J., Squirrell,J., Hollingsworth,P.M., Mackenzie,K., Gibby,M. and Powell,W. (2005) Comparative analysis of population genetic structure in Athyrium distentifolium (Pteridophyta) using AFLPs and SSRs from anonymous and transcribed gene regions. Mol. Ecol., 14, 1681–1695.

Wu,C., Asakawa,S., Shimizu,N., Kawasaki,S. and Yasukochi,Y. (1999) Construction and characterization of bacterial artificial chromosome libraries from the silkworm, Bombyx mori. Mol. Gen. Genet., 261, 698–706.

Xia,Q., Zhou,Z., Lu,C., Cheng,D., Dai,F., Li,B., Zhao,P., Zha,X., Cheng,T. et al. (2004) A draft sequence for the genome of the domesticated silkworm (Bombyx mori). Science, 306, 1937–1940.

Yamamoto,K., Narukawa,J., Kadono-Okuda,K., Nohata,J., Sasanuma,M., Suetsugu,Y., Banno,Y., Fujii,H., Goldsmith,M.R. et al. (2006) Construction of a single nucleotide polymorphism linkage map for the silkworm, Bombyx mori, based on bacterial artificial chromosome end sequences. Genetics, 173, 151–161.

Yamamoto K, Narukawa J, Kadono-Okuda K, Nohata J, Suetsugu Y, Sasanuma M, Sasanuma S, Mita K, Minami H, Shimomura M: Silkworm genome analysis: Construction of an integrated genome database, KAIKObase. Seikagaku 2006, A12627:78.

Yamamoto, T., H. Nagasaki, J. Yonemaru, K. Ebana, M. Nakajima, T. Shibaya, and M. Yano. 2010. Fine definition of the pedigree haplotypes of closely related rice cultivars by means of genome-wide discovery of single-nucleotide polymorphisms. BMC genomics 11, no. 1: 267.

Yang, Z. 2006. Computational molecular evolution. Oxford University Press, USA.

Yang,Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution, 24, 1586-1591.

Yasukochi,Y., Ashakumary,L.A., Baba,K., Yoshido,A. and Sahara,K. (2006) A second-generation integrated map of the silkworm reveals synteny and conserved gene order between

lepidopteran insects. Genetics, 173, 1319–1328.

Yoshido,A., Bando,H., Yasukochi,Y. and Sahara,K. (2005) The Bombyx mori karyotype and the assignment of linkage groups. Genetics, 170, 675–685.

Zdobnov, E.M. & Apweiler, R., 2001. InterProScan-an integration platform for the signature-recognition methods in InterPro. Bioinformatics, 17(9), p.847.

Zerbino, D.R. & Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Research, 18(5), p.821.

Zhang,D.-X. (2004) Lepidopteran microsatellite DNA: redundant but promising. Trends Ecol. Evol., 19, 507–509.

Zhang, J, 2003. Evolution by gene duplication: an update. Trends in Ecology & Evolution, 18(6), pp.292-298.

Zhang, J., Nielsen, R. & Yang, Z., 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Molecular biology and evolution, 22(12), p.2472.

Zhang, L., and W. H. Li. 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Molecular biology and evolution 21, no. 2: 236.

Zou Z, Najar F, Wang Y, Roe B, Jiang H: Pyrosequence analysis of expressed sequence tags for Manduca sexta hemolymph proteins involved in immune responses Insect Biochem Mol Biol 2008, 38:677-682.

Appendices & addenda

Work presented in this thesis can be further explored (including video tutorials) in the internet:

- http://drupal.org/project/gmod_dbsf
- http://drupal.org/project/biosoftware_bench
- http://drupal.org/project/genes4all
- http://gmod.org/wiki/Est2assembly
- http://gmod.org/wiki/Gmod_dbsf
- http://gmod.org/wiki/InsectaCentral

Appendix A – est2assembly user manual

#v0.034 04Jul09 AP

Installation Novice

Dear novice user,

This platform can be used even by novice users, as long as they know basic Unix. A decent tutorial by Lincoln Stein is available here: http://stein.cshl.org/genome_informatics/

Once you are familiar with the basics, you can proceed. If you have any questions at any time, please email me alexie -alpha symbol- butterflybase.org

Please do follow the instructions below in order to have est2assembly running smoothly. I'll now walk you through it. In particular, we are going to install the Perl libraries (modules) which est2assembly relies on. CPAN is the repository where this libraries can be stored and the program cpan can be used to download, test and install them.

For the rest of this manual, where a $ dollar symbol appears, it will represent your shell where you can type commands. It is accessed through the terminal application and comes in various flavours: bash, sh etc. (if you have a choice, I recommend bash as the most user-friendly and least troublesome). In any case, the one already installed will do fine.
Where a > sign appears, then it is commands that follow the previous ones after you press enter (e.g. the cpan shell)

For most of the installation, root (administrator) access is not required but recommended.

Depending on your linux distribution, CPAN is probably installed by default but likely to be severely out of date. Try typing:

```
$ cpan
```

If that does not work, try installing it. In a debian based system (such as the very user friendly Ubuntu) run this as root:

```
$ apt-get install cpan
```

or using sudo (if you are allowed, invoke root priviliges without logging in)

```
$ sudo apt-get install cpan
```

If it is installed, then it is highly recommended that you (or the admin) update it to the current version using administrative priviliges. Keeping CPAN up-to-date is something recommended for any perl program not just this platform.

```
$ sudo cpan
> install Bundle::CPAN
```

```
> exit
> sudo cpan
```

This will take a long time and after re-invoking cpan you will be asked a lot of questions. I hope they are self-explanatory (defaults usually do) or you can figure it out by reading the cpan manual/website as it is beyond the scope of this walk-through. Always ask someone who knows if you are in doubt.

Now that cpan is up-to-date we can continue. We will install the libraries used by est2assembly, then the development branch of BioPerl (as currently the stable one is just too old) and finally invoke the GBrowse installation.

This script will do the all the steps in order. First read the manual (as you should do with any script first**!**)

```
$ perldoc install_perl_modules.pl
```

Then once you understand it, run it as root (recommended) or with the -local installation (if you cannot)

```
$ ./install_perl_modules.pl <options here>
```

The GBrowse installation can be tricky but normally it runs without problems. I have to refer you to their website if you need help: http://gmod.org

Please note that you don't need to have GBrowse installed to run this pipeline. But you do need the development version of BioPerl.

The next (and last) program which must be installed is the emboss package. EMBOSS is a very important package in bioinformatics and if you haven't installed it yet, I hope you find it useful beyond this pipeline.

Installation must be done as an administrator or with admin priviliges (sudo). In Ubuntu (or other debian) simply
$ sudo apt-get emboss

Other systems, please refer to the website (http://emboss.org)

Last is the convenient but not necessary step of adding the path of est2assembly to your "executable path" (the $PATH environmental variable) or copying all the perl scripts to your bin.

To put the directories in the path you can execute the following command from the directory where this file resides.
$ export PATH=$PATH:$PWD/preprocess:$PWD/

The $PATH is your old PATH variable which is propagated, the $PWD is a special variable which contains your current directory. Depending on your shell, if you press tab it will autocomplete to the current directory.

try

```
$ echo $PWD
```

and you will get the real full path of your current dir.

If you want this information to be stored when you start your computer, first get the line with $PWD replaced to the real full path.
Then in your startup file for your shell (for bash it's ~/.bashrc) where ~ is a special character/shortcut for your home. To install it for all users you can put it at /etc/bash.bashrc

e.g.

```
$ pico ~/.bashrc
```

or

```
$ vi ~/.bashrc
```

Alternatively, if you already have a ~/bin/ directory and it's in your path, you can run

```
$ install_into_bin.sh
```

To copy the files to your ~/bin/ (or other directory: remember to read the manual with perldoc).

Finally you have to install some external programs. Please read the INSTALL file for the relevant websites. We include those we are allowed by their license to.

That's it! For expert users this is probably trivial and will not take more than a few seconds but I hope this short manual is sufficient in detail for new administrators.

If you think it can be improved, please do send suggestions to alexie -alpha symbol- butterflybase.org. Even better, you can correct/expand it and email it and I'll include it in the next release (along with a big thank you).

Installation Expert

Dear expert user,

I hope the platform is self-explanatory to you, especially after reading this manual. The perl libraries can be installed by reading the Cpan_packages_needed.pm file to see which are needed, then also install the -dev branch of BioPerl and if you wish GBrowse.

Or you can do all of the above using the install_perl_modules.pl script.

You can put the scripts in your path or copy them to your bin. Please do keep the lib TrimByWindow properly connected or install it site-wise. Also the .config files are not needed as they can be specified by the command line interface but if you do want to avoid typing, please put them in the same

directory as the relevant scripts ($RealBin/ is used to find them).

That's it! If you think it the manual can be improved, please do send suggestions to alexie -alpha symbol- butterflybase.org. Even better, you can correct/expand it and email it and I'll include it in the next release (along with a big thank you).

Manual conventions

For the rest of the manual, I would like to point out to the following conventions:

- Dollar symbol $ denotes the command line.
- The brackets <> symbolise an option which can be replaced with text of your choosing. Don't include <> in the command line unless otherwise specified.
- The perldoc command ($ perldoc <perlscript>) can be used to see the documentation of any script. Limited usage information can be gained by running the script without any options.
- An option can be activated with the option's full name or an abbreviation that is unique versus all the other available options in the script. Alternatives are shown separated by the horizontal bar | so that if two options/switches, d|debug and dir, exist then using -d will call for -debug and -di will call for -dir.
- The options sometimes have :i, :f or :s. :i denotes that an integer is expected for this option, :f expects a float (i.e. decimal) and :s it expects a string, i.e. a text. If nothing is given, then the option does not accept values but is a switch.
- If you include spaces in your values, then you must quote the value (e.g. in " "). Also if an option has the {,} denotion, it means that more than one value can be given to this option, either through specifying multiple times the switch or putting the values one after the other. See $ perldoc Getopt::Long for detailed information for this system.

e.g.

```
$ ic_loaddata.pl
Please provide caf/p4e/annotation files or a config directory
Usage:
            'c|caffiles:s{0,}'              => CAF assembly files
            'p|p4efiles:s{0,}'              => prot4EST files.
            'b|blast|annot8rfiles:s{0,}'        => includes blasts, snps, annot8r,ipr etc
            'host:s'                        => DBHOST
            'port:i'                        => DBPORT
            'prefix:s'                      => Prefix for psql database (def "ic_")
            'create|d|drop|recrete'         => Drop and recreate psql database,
            'a|config|gbrowse_config:s'     =>  directory to gbrowse.conf which has include files
            'species:s{,}'                  => Species list
            'types:s{,}'                    => caf, orf or p4e
```

```
'debug'                         => Don't delete load tmp files
```

So the -caffiles can be given as -c and the -create can be given as -cr or as -d. Caffiles, p4efiles and annot8rfiles can have multiple values passed to them (with a minimum of `0`, i.e. none). The -debug and -create options don't expect any values and if one is given, it will be ignored.

Downloading mRNA from GenBank

I have included a simple script to download EST, mRNA or even protein sequences from EBI (European Bioinformatics Institute), part of the GenBank consortium.

```
$ srsdownload.pl -s "Drosophila melanogaster" -m cDNA
```

will get all the cDNA (mRNA+EST) from D.mel. It will take some time as the script waits every 1000 requests so that their server doesn't complain.

In a transcriptome assembly we include these sequences (as Sanger technology) in order to improve the quality of the assembly (especially full length cDNA). MIRA allows the use of strain information so you can use

```
$ mira_strain.pl
```

to produce a strain file for GenBank data to concatanate into the final strain file.

Running a BaseCall for Sanger capillary sequencing

The script which drives the basecallling for sanger capillary sequencing is called (surprisingly) process_sanger_trace.pl.

As always, read the manual...

```
$ perldoc process_sanger_trace.pl
```

You can find where it is installed (if you didn't install it yourself)

```
$ which process_sanger_trace.pl
```

and maybe using

```
$ ls -l <output of above>
```

To see if it is a symbolic link to another directory. The configuration file should be in the same directory as the script.

We find that the newest (beta) version of phred is quite good (as good or better than the KB Basecaller if you use an ABI sequencer) but which basecaller you wish to use, is up to you. Note that the *Trimbywindow* routine can only be used with sequences which have integrated the KB basecalling in the .abi file.

The configuration file, residing in the same directory as the script, can be configured to recognise your sequence naming scheme (hopefully you used a scheme and not ad-hoc named each sequence...). You will do know some basics about Perl regular expressions (regex). E.g. see http://www.cs.tut.fi/~jkorpela/perl/regexp.html

Preprocessing your data

It is essential to preprocess your sequence data before trying to assemble it. You don't want to allow false local alignments in the assembly as these will not only result in misassemblies but also a high percentage of read rejection.

You probably want to mask/remove adaptor sequences added during the library construction, identify any vector sequence (if sequencing with capillaries) and common contaminants. Identifying and removing known repeats and transposons is also highly recommended as it will result in having 'cleaner' data for the assembler to process. If you are interested in transposons, then by having the above identification, you can annotate them separately.

The script preprocessest.pl (in the folder est_assembly) does all this for you in one go. It can be slow but it is thorough. Most importantly, it keeps a log file to what happened in each sequence so we can check (a database version is in my "to-do" list) As always, read the manual

```
$ perldoc preprocessest.pl
```

- It has a config file to specify database variables if you are using it often. But you can override them from the command line. If you don't want to use a specific database run at all (e.g. there is no adaptor sequence in your data (?!)) then please give the relevant -no option (e.g. -noadaptor).
- You can choose to keep the intermediate files, archive them in a tar.bz2 file or delete them. They do take an awful lot of space but they can give you valuable information about your data.
- The files useful to the assembler are the ones which have the word cleaned in them. They have the full sequence masked with Xs ("just in case" TM) and an XML files tells the MIRA assembler which parts to use. If you want, you can use the original sequence to give to the assembler (if

you think there is over-masking and wish to take advantage of MIRA's option to identify such regions). Use this command

```
$ trim_fasta_all.pl -inv -id <yourcleaned file> <your original file>
```

This will produce two files .trim and .discard. The .trim you wish to use, the discard you can look at and see what preprocessest has rejected (there are also logfiles in the interemediate files which will show you the same thing).

A second set of files is produced for Newbler or any other program that doesn't accept XML files. These have the bases removed. You can use them, for example to do BLASTs with or against the unassembled reads.

General tip: For BLAST you should generally make sure that there are no IUPAC codes in your data (some versions convert them to X). The script cleaniupac.pl converts a IUPAC DNA code to a ATCG code (see its perldoc).

NB: I recommend processing each technology/input dataset in a separate directory. If you have datasets from same technology but generated with different methods (e.g. vectors) or you are using GenBank derived data, then you should concatanate data from the same technology before the assembler. use the cat command

```
$ cat [fastafile1 fastafile2 ... fastafileN] > outfasta
```

For XML files, you can concatanate them with the merge_trace_xmls.pl script. As always, see its perldoc for how to use it.

Feel free to use any other subscripts in est_process for whatever you want. They all should have a perldoc describing what they do and how to use them.

Walkthrough

Let's try an example. One based on Sanger as 454 will take too long and the output is no different.
So, do you remember the Drosophila melanogaster ESTs we downloaded with srsdownload.pl ? Oh, you didn't do that... That's OK, let's do it together and see how it works, but we will download something smaller, such as ***Heliconius melpomene*** ESTs (a butterfly). First create a directory to download and process them. I'm assuimg you are already into a 'parent working directory' where you store your work.

```
$ mkdir melpo_cdna
$ cd melpo_cdna
$ srsdownload.pl -m cDNA -s "Heliconius melpomene"
$ ls
```

It should take about 90 seconds. We will have a file 34740.cDNA.fasta contained the sequences. The first number is the NCBI taxonomy ID for our species (using IDs is safer, avoids misspellings). These should be

preprocessed (if they are in GenBank) but let's do it anyway. Because we don't know what adaptor they used (and we don't have time to check with GenBank/publications) we will use the -noadaptor option and hopefully they (actually me some years ago) got that part right.

Let's assume you don't have a config file setup yet, so everything will be in the command line. You will obviously have to correct the paths to match your system

This will remind us the options

```
$ preprocessest.pl
```

```
$  time preprocessest.pl -noadapt -de /db/blastdb/ecoli_pseudomonas.fsa.masked.nr -dr /db/blastdb
/rDNA_nt_inv.fsa_nr -dm /db/blastdb/mito_aa_inv_80.fsa.trim -rl /db/blastdb/repeats
/insecta_extra.repeats_nr95 -strain GenBank -thread 4 -deletetmp -backuptmp -archivelog -tech
sanger -p Melpo_genbank 34740.cDNA.fasta
```

It will use four threads for BLAST and SSAHA (so make sure you have at least 4 CPUs available or reduce the number; you can also increase it). It will also produce a strain file with GenBank as the id. Files will have the prefix Melpo_genbank.

The time command in the beginning is completely optional. It will tell (at the end) how long it took (real time i.e. wall-clock time) and the actual CPU time (called 'user' i.e. don't include the time waiting for the hard disk to read/write). See also time -v

This will eventually (5m59sec on my computer) produce the assembler input files, a set of logfiles archived in a tarball and backup the intermediate files before deleting them. That's it!

I put the output in the test folder of est_process so you can verify against your version.

If you want to reduce the command line next time, simply edit the config file found in the same directory as the script, or produce a new one (if you want to keep track for different projects) and provide it with the -config option. The layout is a simple text file with these variables:

```
adaptor_db="/db/custom/454adaptors.fasta"
mini_adaptor_db="/db/custom/454restr_sites.fasta"
ecoli_db="/db/custom/ecoli_pseudomonas.fsa.masked.nr"
rdna_db="/db/custom/rDNA_nt_inv.fsa_nr"
mito_db="/db/custom/mito_aa_inv_80.fsa"
repeat_lib="/db/blastdb/repeats/insecta_extra.repeats_nr95"
```

TIP:
If you are using 454 data, then you must include an adaptor library. Ask your sequence center/provider. If you don't... well... you won't enjoy it the results of the assembly...

Running an Sanger and/or 454 assembly with parameterization

The most challenging process in assemblies, is knowning which parameters to use. Newbler and other black boxes, allow you to go for the default settings and get /an/ assembly but not necessarily /the/ assembly you want.

The script parameterize_assembly.pl helps you to do this but should not be used as a black box either. Please read the MIRA manual so you know which parameters you wish to tweak. Then the config file can edited to allow you to produce runs with the different settings and compare them with certain statistics (please see our est2assembly paper). I do provide a config file for the brave ones amongst you.

Configuration file

The format of the config file is simple. If there is # character, then this line is ignored. then we have the basic settings which are always appended if the relevant technlogy is used:

`mira_settings_basic` -> Always used
`mira_settings_sanger` -> For sanger
`mira_settings_454` -> For 454 data.

Then we have the parameters which you wish to test, one per line:
They start with the word "parameter:" and come in pairs of key and value. The line is comma delimited (except parameter which is colon delimited)

e.g.

```
parameter:
settings,-SKIM:rt=10,settings_454,-AS:ardct=3,settings_sanger,-AS:ardct=3,assembler,MIRA,description,$
Repeat coverage multiplier increased to 3 repeat threshold defaults to 10
```

translates to:

- GENERAL_SETTINGS in mira will include -SKIM:rt=10
- SANGER_SETTINS will be -AS:ardct=3
- the assembler is called MIRA
- the description in the log file is "Repeat coverage multiplier increased to 3 repeat threshold defaults to 10". The special variable $runquality will be replaced with actual run number in the log file so just add it in.

We use reference databases to benchmark each assembly run. If you wish to use friendly names in the log output, then add lines like these:

```
database:Agambiae_ref_vbase_uniprot_100, Anopheles RefSeq VectorBase and Uniprot (nr100)
```

Reference databases

We recommend you use protein databases from the same organism you're sequencing or at least closely related (but even same phylum will work fine). For benchmarking it is not terribly important as we are interested in the relative performance rather than the absolute one. However, if your reference database is not very good (i.e. too distant) then the benchmarking will not have sufficient data to make a comparison robust.

Output

When an assembly is finished, a comma separated file can be loaded in a spreadsheet program to determine which assembly is optimal for you. Choosing an assembly is subjective (please read our paper) and depend on the aims of your project. We tend to go for those that maximise coverage of a reference proteome and maximise the number of reads even if the annotation redundancy is high. Annotation redundancy is essentially how many multiple hits your assembly has versus the reference database (totalled in both directions).

Walkthrough with *H. melpomene*

Now we can continue with our walkthrough of analysing the *H. melpomene* Sanger sequences downloaded from GenBank. The following output exists in the test folder of parameterize_assembly so you can check the output against what you will try yourself now.

make a directory (mkdir) where the assemblies will run and make a note of where your Melpo_genbank_Sanger.cleaned.fasta.x file is. For me it is at ../../../est_process/test/melpo_cdna/ relative to my current working directory.

```
$ time  parameterize_assembly.pl -f1 ../../../est_process/test/melpo_cdna
/Melpo_genbank_Sanger.cleaned.fasta.x -t1 sanger -p Melpomene_assembly_genbank  -accurate -config
parameterize_assembly_dbest.conf -log -newbler ../../../est_process/test/melpo_cdna
/Melpo_genbank_Sanger.cleaned.fasta.x.in_newbler -db /db/blastdb/silkpro_V2_ref_db_uniprot_100
--thread 6
```

- This command will launch the parameterization of the assembly using the FASTA file we produced last time.
- It will automatically find the qual and xml files if they are simple extensions of the fasta name +.qual +.xml respectively.
- We specify that the input file f1 is from Sanger capillary technology (via t1)
- We give the project a name Melpomene_assembly_genbank and request an accurate assembly.
- We will use a default config present in the distribution that I use for dbEST (feel free to customize).
- We are asking for a log file to be kept
- Also to run newbler using the fasta file that has the masked sequence removed. Newbler will also find the quality file by itself.
- Then we are specifying which database we will use for benchmarking the different parameter (the Silkmoth proteins from RefSeq + Uniprot + SilkDb made non-redundant at 100% identity). You can include multiple

-db options, one after the other.
- Finally we are asking for 6 parallel threads (CPUs).

Tip while you wait:
You might as well go for lunch now, as this will really take a while. If you are running this from a computer connected to a server and you're afraid of losing your connection, then add screen in front (if it is installed on your server). This will put the command on a separate process in the background. You can later fetch it via screen -r

```
$ screen time  parameterize_assembly.pl -f1 ../../../est_process/test/melpo_cdna
/Melpo_genbank_Sanger.cleaned.fasta.x -t1 sanger -p Melpomene_assembly_genbank  -accurate -log
-config parameterize_assembly_dbest.conf  -newbler ../../../est_process/test/melpo_cdna
/Melpo_genbank_Sanger.cleaned.fasta.x.in_newbler -db /db/blastdb/silkpro_V2_ref_db_uniprot_100
>Ctl A then Ctl D
```

The last key combination will detach the screen which carries on the background. Later, to connect (if it is still running):

```
$ screen -r
```

If it has finished then the screen closes automatically and you cannot see any output (including errors). This can be ok if you don't expect any. If you might then use screen first to create the 'pseudo-terminal' and then launch the command;

```
$ screen
$ time  parameterize_assembly.pl -f1 ../../../est_process/test/melpo_cdna
/Melpo_genbank_Sanger.cleaned.fasta.x -t1 sanger -p Melpomene_assembly_genbank  -accurate -log
-config parameterize_assembly_dbest.conf  -newbler ../../../est_process/test/melpo_cdna
/Melpo_genbank_Sanger.cleaned.fasta.x.in_newbler -db /db/blastdb/silkpro_V2_ref_db_uniprot_100
> Ctl-A then Ctl-D
```

You can use screen -r to go back.

Next time you use this program, you might want to include a 454 dataset. simply give -t1 as 454, or if you use both sanger and 454 then keep -t1 sanger and add -t2 454 -f2 <your 454 filename from preprocess>

Continuing with our walkthrough:
Once the assemblies have all finished, we can look at the output:

```
$ less Melpomene_assembly_genbank.accurate.blast_analysis.csv
```

or you can use a spreadsheet program.
A summary of Melpomene_assembly_genbank.accurate.blast_analysis.csv is available as Melpomene_assembly_genbank.accurate.blast_analysis.total.csv. It has only the lines with Total and is applicable if you used more than one reference proteome.

How you choose the assembly is subjective (see above) but I think

Melpomene_assembly_genbank.accurate.2 is pretty good (if you use the same version of MIRA you should get the same result). You can also see that Newbler does not perform very well (hopefully someone in 454 Life Sciences reads this!).

I personally either delete all the runs I do not want

```
$ rm -rf <directory>
```

careful this is a very powerful command : it does not ask for confirmations and it deletes directory structures: in Linux if you delete smth it's gone! (but see *libtrash*).

Did I mention you should be careful with rm -rf ?

the other option I use when I want to keep the runs (just in case) and there is no problem with disk space is to make a soft-link, i.e. a shortcut e.g.

```
$ ln -s Melpomene_assembly_genbank.accurate.2 chosen
```
or if you want the result directory:
```
$ ln -s Melpomene_assembly_genbank.accurate.2/Melpomene_assembly_genbank.accurate.2_d_results/
chosen
```

then you know that chosen is your assembly of choice and you don't have to remember which one it was.

Tip:
An assembly is not a clustering process. Please make sure you understand this principle while working with EST projects.

Assembly annotation

BLAST annotation

Once an assembly is chosen, it should be annotated so you data mine it. First you have decide what kind of databases you wish to annotate with. Then make the relevant BLAST files.

Also you might want to consider to add Enzyme Classification, Gene Ontology and KEGG pathway terms using a similarity to an annotated protein. In order to do this, you will need to predict a protein for each contig.

Walkthrough for BLAST

Let's continue with annotating our Melpomene data. I put the examples in the analyze/test directory.

You have two options: running blastall directly (make sure it is installed) or using the analyse_assembly.pl script. The latter option does some extra work

(and takes extra time) as it calculates the b.p. coverage and other stats (same method as in the parameterization) and has the option to extract the parts of the query which have a hit (the -extract option) in case you want to look at it.

Let's produce an analysis using the known proteins (uniref100 from UniProt):

let's make a local link to the input file which will also be renamed to smth more friendly (to make future commands shorter). (in your system the input file might have a different path**!**)

```
$ ln -s ../../../../parameterize_assembly/test/melpo_cdna_assembly/chosen
/Melpomene_assembly_genbank.accurate.2_out.unpadded.fasta Melpo_assembly.fsa
$ analyse_assembly.pl -i Melpo_assembly.fsa -db "/db/blastdb/uniref100.fasta" -t 10 -ce 1e-5 -cs
80 -extract
```

Note that you have to give your own -db argument for uniref100 (also the quotes are not necessary). Also make sure you have enough CPUs for the -t argument (threads). BLASTx versus Uniref100 will take a (very) long time so feel free to use a smaller database if you prefer (uniref90, uniref50 or even the reference proteome BLASTx you have already completed during the assembly). My computer took 207m36.065s for the above command.

The -cs and -ce arguments give cutoffs for score and e-value (which do not correlate with each other). It is quite stringent but a bit-score lower than 80 will produce false alignments. Lower it if wish. The BLAST report itself is actually run with lower cut-offs (1e-1) The downside of this is very large BLAST reports (0.6 Gb for the above one) but it does allow investigation of lower cut-offs. The HASH used for the analysis is read with the specified options above. This means you can use *analyze_blast.pl* change the cut-offs without having to re-run the BLAST (or re-write the hash; use the -uh option).

e.g.

```
$ analyze_blast.pl -cs 60 -ce 1e-1 -uh Melpo_assembly.fsa.x.uniref100.fasta.refblast
```

The .x in Melpo_assembly.fsa.x.uniref100.fasta.refblast appears because the input file converted all IUPAC codes to a valid nucleotide A,T,C,G as BLAST is notorious for not liking them (I believe maybe fixed in a new version, but better to be sure). You can switch off this behaviour with the -noclean argument of `analyse_assembly.pl`

Also note that the analysis of the BLAST is not necessary for the protein annotation below. It does provide important information about your assembly though.

NB. analyse_assembly.pl compresses the BLAST report after it is finished with it (to save space). So if you need, please uncompress it with bunzip2 again.

```
$ bunzip2 Melpomene_assembly_genbank.accurate.2_out.unpadded.fasta.x.uniref100.fasta.refblast.bz2
```

or

```
$ bunzip2 -k
Melpomene_assembly_genbank.accurate.2_out.unpadded.fasta.x.uniref100.fasta.refblast.bz2
```

to keep both the BZ2 and BLAST files.

Let's also create BLASTs against a mitochondrial and rDNA database (one for Insects is included in distribution). No need to analyse it so just use the normal blastall executable.

```
$ blastall -i Melpo_assembly.fsa -d "/db/blastdb/mito_aa_inv_80.fsa.trim" -Q 5 -o mito.bls -p
blastx -a 5
$ blastall -i Melpo_assembly.fsa -d "/db/blastdb/rDNA_nt_inv.fsa_nr" -o rna.bls -p blastn -a 5
```

The warning about the -threads option applies here too for the -a option.

Protein prediction

Please do note that not all contigs will be from coding regions (an assumption frequently made and an honest mistake) so forcing a protein prediction may result in an utter unrealistic one. Nonetheless, a protein prediction for the whole assembly is recommended and we use a program called prot4EST. A version I have made a few tweaks on is included in this distribution.

The program (you should also read its paper...) relies on two main approaches that work and two that generally don't : similarity to known protein, ESTScan, Decoder, longest ORF. A contig is first tried with the first two and if these fail it tries with the other two. ESTScan requires a organism-wide codon-usage table (the concept of which is arguable to be realistic as loci can utilise different codon usage but let's ignore this now). How we produce it, we will talk about once we cover the similarity step.

We suggest you run a BLASTx versus Uniref100 from Uniprot. The full blast report must be provided in the configuration file of prot4EST (see the prot4est directory). Using Uniref100 will make it more likely you pick a similar hit than using a reference proteome (as you did before) but it will be much much slower (because parsing such reports and tiling the high scoring pairs (HSPs) in a hit is rather slow). Any proteins predicted with this method are almost certainly real. Prot4EST uses an internal cut-off score so feel free to use whatever cut-off you think is real. Once you have this BLAST report you can use it to build a codon and HMM model for ESTScan.

- The easy way to build one is if you had a lot CDS data in GenBank for your species, but that is probably unlikely if you are trying to predict proteins from ESTs. If you have used the srsdownload.pl option, then you probably have already included those CDS in the assembly.
- There is the 'hard' way which is also more reliable but also made easy due to a script I provide, it pretty much automates the whole procedure.
  - We will build a set of reliable proteins using the BLASTx report from above; even if they are partial they will be enough to build a model of codon usage and a HMM for ESTScan.
  - The HMM model is build using a reference proteome which is then backtranslated using the codon usage table created.
  - So in effect, we create fake CDS regions based on known proteins using your species' codon usage model (as determine through your

assembly).

See the options:

```
$ ic_build_codons_model.pl
```

There will be two files of importance: the .cod is the codon usage table and .smat is the HMM model for ESTScan. Edit the prot4EST file to include these two.

To speed things up, you also need to do a BLASTn versus rDNA genes and a BLASTx versus mitochondrial genes. Be careful; when you do the BLASTx versus mitochondrial genes, you should make sure you provide the correct codon table number using the -Q option of blastall (e.g. -Q 5 for insects). Name these output files rna.bls and mito.bls respectively and let them reside on the working directory where you will run prot4EST from (and where the config file resides). See the walkthrough about BLAST above.

Walkthrough for protein prediction

So let's predict proteins for our Melpomene data: I will use the prot4ESt/test /melpo_p4e directory to store my results so you can compare.

Create symlinks (paths may differ; don't forget the dot as target at the end):

```
$ ln -s ../../../analyze/test/melpo_annotation/rna.bls ../../../analyze/test/melpo_annotation
/mito.bls ../../../analyze/test/melpo_annotation/Melpo_assembly.fsa.x.uniref100.fasta.refblast
../../../analyze/test/melpo_annotation/Melpo_assembly.fsa.x <path to
assembly>/Melpo_assembly.fsa.qual  .
```

Make a copy the empty p4e.config file

```
$ cp ../../p4e.config .
```

Make a dir for our codon modelling and build it

```
$ mkdir codons
$ cd codons
$ ic_build_codons_model.pl -b Melpo_assembly.fsa.x.uniref100.fasta.refblast -i Melpo_assembly.fsa
-a "/db/blastdb/silkworm/silkpro_ref_uniprot.clean_nr90" -s "Heliconius melpomene"
```

Finally you need to create a special directory proteins.x with FASTA sequence and quality where the files are split (using splitfasta.pl) for each contig (from the assembly). Please use the unpadded files and make sure the sequence has no IUPAC codes (prot4EST will break):

using the linked files from above:
$ splitfasta.pl -i Melpo_assembly.fsa -suffix seq -dir proteins.x
$ splitfasta.pl -i Melpo_assembly.fsa.qual -suffix qlt -dir proteins.x

Once you have all the files ready, edit p4e.config and point to the correct files. Then run it with this command (from the directory containing the p4e.config).

```
$ prot4EST_2.3-AP.pl 1
```

The 1 option is to make the process automatic. You might want to use screen again, as this will take also a long time (due to parsing the Uniref100 BLAST file using BioPerl). It took 30' on my computer.

Proper naming

This step can be run at any time until now, but it should be run before the use of annot8r next.

As we are going to database our data, it is important to provide a standard approach to naming them. The naming scheme recommended (and a script provided for) is this:

Generally it looks like this
<a unique database ID of your choice><species ID><Assembly identifier><data type><serial number>

Specifically, we are going to use:

- For an **assembly contig**: <two letters for db id><NCBI Taxid number><assembly run where Aa is 1 Ab 2 etc><E for expressed><con for contig><serial number>
- For a **peptide/ORF**: <two letters for db id><NCBI Taxid number><assembly run where Aa is 1 Ab 2 etc><A for automated prediction><pep for peptide or orf for open reading frame><serial number>
- For a **marker**: <two letters for db id><NCBI Taxid number><assembly run where Aa is 1 Ab 2 etc><M for marker><snp for SNP, ssr for SSR etc><serial number>

The DB id should be the same in your Lab. We use IC (for InsectaCentral). The assembly run is meant to be used when new assemblies are created (by adding new data or updating something) not by the assembly parameter set chosen.

This naming scheme ensures that each object is completely unique and can have exist even after an assembly is retired. If needed, link tables can provide such historical links.
Please also note, that the protein/orf serial id follows each other but not the contig id. This enables us to store multiple protein predictions for the same contig. The prot_main.psql.named link table in the prot4EST output directory provides the links between them (and is used by subsequent steps such as SNP alignment).

So let's name our assembly file (in CAF format), the FASTA files (the unpadded sequence/quality files and the padded quality for the SNP analysis later), the BLAST reports and the protein results. All except the later involve a simple

pattern search.

Renaming the prot4EST includes a special feature not available in prot4EST 2 but will become available in prot4EST 3 (I don't know the release date, so please don't ask me...). As you probably know (because you read the prot4EST paper), it works by predicting and correcting the protein translation (due to frameshifts). As a result the open reading frame is not simply a sublocation within the contig. For that reason ic_create_naming.pl will try to find the correct sequence of the ORF by trying to align the protein with the contig using fasty34 (by Pearson) and if that fails by using backtransambig (from EMBOSS) and correcting the ambiguous nucleotides where possible using the contig. This can take quite a long time for 454 assemblies and I know it is not the best implementation but as the next version of prot4EST is due to implement this, I saw no reason taking it apart and doing it myself.

We will use DM for database name (for DeMo). It will be the first assembly we named and made public, so we don't have to set -assembly 1 (it's the default). If we know the NCBI taxid we can give it in the -species option directly, instead of writing "Heliconius melpomene".

The output files can be specified, the default is the original file +.named.

Go to the directory where the assembly is and rename the Assembly

```
$ ic_create_naming.pl -t caf -d DM -s "Heliconius melpomene"
Melpomene_assembly_genbank.accurate.2_out.caf
```

Rename the FASTAs

```
$ ic_create_naming.pl -t fasta -d DM -s "Heliconius melpomene"
Melpomene_assembly_genbank.accurate.2_out.unpadded.fasta
Melpomene_assembly_genbank.accurate.2_out.unpadded.fasta.x
Melpomene_assembly_genbank.accurate.2_out.unpadded.fasta.qual
Melpomene_assembly_genbank.accurate.2_out.padded.fasta.qual
```

Go where the BLASTs are. Rename the BLASTs.

```
$ ic_create_naming.pl -t blast -d DM -s "Heliconius melpomene"
Melpomene_assembly_genbank.accurate.2_out.unpadded.fasta.x.uniref100.fasta.refblast
```

Go to where you run prot4EST from. Rename the prot4EST output. The -type is p4e and the input will be the output directory where the files have been produced (on my computer this took about one minute)

```
$ ic_create_naming.pl -t p4e -d DM -s "Heliconius melpomene" output/
```

Again you will see a fasta file with an X: translations_xtn.fsa.X, translations_xtn.fsa.X.named. Prot4EST sometimes produces ambiguity codes (e.g. B) which are not recognised by many software. These amino acids have been converted to X (i.e. unknown).

you can rename your files to match the header:

```
$ head translations_xtn.fsa.X.named
$ mv translations_xtn.fsa.X.named  DM34740AaApep.fsa
```

Annotate with EC/GO/KEGG and walkthrough for annot8r

Now we can use the proteins predicted by p4e to BLAST versus known proteins which have an EC, GO or KEGG term. The relevant postgres databases are included in the annot8r directory, along with a hacked copy of annot8r (again, you should read their paper and use their software to understand what it does).

If you want to create new Uniprot FASTA files or new psql databases, then you need to use the original annot8r software. I provide, only for convinience, the psql and BLAST databases that were current on 01Jun2009.

Go to the annot8r directory. Extract the psql files

```
$ bunzip2 *psql.bz2
```

Assuming your postgres (psql) installation is working correctly, if you have more than one postgres cluster (unlikely if you are a novice user)
$ pg_lsclusters
And make note of the port number you wish to use. You will have to provide it as -dbport for annot8r or -p for psql. Likewise for host (with -dbhost or -h respectively.)

Load your psqls

```
$ psql < a8r_ecbase.psql
$ psql < a8r_gobase.psql
$ psql < a8r_keggbase.psql
```

And three new databases, a8r_ecbase, a8r_gobase and a8r_keggbase will appear in your psql cluster.

Overview of how to use it is, as always, available with perldoc:

```
$ perldoc annot8rAP-gff.pl
```

Essentially, you are taking a protein FASTA file, BLASTing it against known proteins with KEGG, EC, GO annotation and based on a particular cut-off you can assign KEGG, EC or GO terms using the IEA (Inferred Electronic Annotation) code.

Annot8rAP-gff.pl needs as input BLAST files. Produce them with the method of your choice using each of the databases (which you will need to uncompress first)

In the annot8r directory:

```
$ tar -xjf uniprot_GO.fsa.tar.bz2
$ tar -xjf uniprot_KEGG.fsa.tar.bz2
$ tar -xjf uniprot_EC.fsa.tar.bz2
```

BLASTP them with your protein translations,
e.g.

```
$ blastall -i DM34740AaApep.fsa -o DM34740AaApep_vs_EC.blastp -e 1e-5 -d "/db/blastdb
/uniprot_EC.fsa" -a 2 -p blastp -b 30 -v 30 2>errors
$ blastall -i DM34740AaApep.fsa -o DM34740AaApep_vs_KEGG.blastp  -e 1e-5 -d "/db/blastdb
/uniprot_KEGG.fsa" -a 2 -p blastp -b 30 -v 30 2>errors
$ blastall -i DM34740AaApep.fsa -o DM34740AaApep_vs_GO.blastp  -e 1e-5 -d "/db/blastdb
/uniprot_GO.fsa" -a 2 -p blastp -b 30 -v 30 2>errors
```

The 2>errors will redirect any errors to the errors file. Feel free to check it between BLASTs. Make sure the -a is optimized for your system and the -d argument is correct. Expect the BLAST vs GO to take quite some time (hours). We will also estimate some parameter/statistics for each protein with the annot8r_physprop.pl script

Then simply

```
$ annot8rAP-gff.pl DM34740AaApep_vs_EC.blastp DM34740AaApep_vs_KEGG.blastp
DM34740AaApep_vs_GO.blastp
$ annot8r_physprop.pl DM34740AaApep.fsa
```

Annot8r will ask the postgres databases you made in the last step in order to decode the hits and then produce an annotated file ready to be converted to GFF (see next chapter).

Annotate with InterProScan

Please see the LSF section and the InterProScan manual as how to use someone else's program is not the aim of this manual. InterProScan can be tricky to set up and takes an awful long time to run (up to 10 min per protein) so please use only if you are an advanced user. We highly advise the use of PC-Farm but I must caution you: please work with your system administrator as iprscan can IO starve the servers (and therefore crash them).

Once you produced an iprscan result in raw format (it must be in raw) then you can give it as argument to the GFF creation program later on (along with the other annot8r output files).

Annotate with SNPs

Single Nucleotide Polymorphisms are very important for a variety of reasons (which I will not delve into). One way a SNP can shown is as a polymorphic base in the consensus (using the IUPAC nucleic acid codes). The assembler MIRA is producing such SNP output and converting it to 'Tags' which the caf2gff script later on will allow you to extract this information. For many applications, however, a more stringent search may be required, and therefore I provide a method to detect SNPs de-novo from the assembly (I call them high-quality predicted SNP): the polymorphic nucleotide has to exist in at least two or three reads and have an certain number of invariable positions up- and down-stream of the SNP (called padding). The script that automates the process is called ic_create_snps.pl.

Walkthrough for SNPs

Go to the directory where the assembly resides, (especially the CAF file) and use the following command.

NB: Due to the programs used it will take quite a long time, especially with 454 assemblies. There is a chance that I will rewrite the script to not make use of SEAN and be a bit faster, but at the moment this is not a priority.
NB2: You will probably wish to use the .named files in order to basecall the SNPs, but you don't have to. It can be done in the GFF stage

```
$ cd parameterize_assembly/test/melpo_cdna_assembly/chosen
$ ic_create_snps.pl -caf Melpomene_assembly_genbank.accurate.2_out.caf.named -cov 2 -padd 20
```

This will identify SNPs if the minor allele occurs at least two times and there are 20 bp invariable up and downstream. The script will use convert_project from MIRA to create a FASTA file of all the ESTs that exist in the assembly (you can also provide the original EST FASTA file if you don't want to wait, the script looks for the <caf filename> + the .fasta suffix), and then it creates an ACE file (to use with SEAN). It splits the ACE file to one ace per contig and runs SEAN on each one.

A simple text file called "snps" containing all the SNPs for each contig will be printed out but this will not be very informative due to the method SEAN uses to parse the files. You will need to make the GFF file, which brings us to the next section.

Creating GFF files

In order to distribute or use the information you just created in a more standard format we will make use of the GFF file format (read http://gmod.org/wiki/GFF). est2assembly contains a number of tools to create valid GFFs for Chado or Bio::DB::SeqFeature format from all the data types you have generated so far.

So let's look at the gffcreation folder and our Melpomene test

go to gffcreation/test/melpo_gffs and we will create a GFF for each file. By not providing the -o (output) option, all GFFs will be created in the current dir and will be named as: <input file> + .gff (and also +.biofeature for BioFeature compatible files).

Creating a GFF of the **assembly** is straightforward and we support both contig-read and read-contig formats:

```
$ ic_caf2gff.pl -e -org "H.melpomene" -biofeature "../../../parameterize_assembly
/test/melpo_cdna_assembly/chosen/Melpomene_assembly_genbank.accurate.2_out.caf.named"
```

Likewise, the GFF for **BLAST** is straightforward (but for Uniref100 in 454 assemblies, it will take a long time - due to opting for the robust BioPerl processing method...).

```
$ ic_blast2gff.pl -cs 80 -o "H.melpomene" -biofeature -l 10 "../../../analyze
/test/melpo_annotation/Melpo_assembly.fsa.x.uniref100.fasta.refblast"
```

The **Protein Prediction** GFF is straightforward to make. You need the output

of prot4EST and the ic_create_naming.pl.

```
$ ic_p4e2gff.pl -main "../../../prot4EST/test/melpo_p4e/output/prot_main.psql.named"  -blast
"../../../prot4EST/test/melpo_p4e/output/prot_hsps.psql.named" -pt "../../../prot4EST
/test/melpo_p4e/output/DM34740AaApep.fsa" -nt "../../../prot4EST/test/melpo_p4e/output
/orf.all.fsa" -e -o "H.melpomene" -biofeature
```

For the SEAN output, we need to give the directory where the splitted ACE files reside and the file containing the quality values (which must be padded, i.e. align to the assembly). In order to identify if a **SNP** is coding and synonymous/non_synonymous, we also need some of the protein prediction files.

```
$ ic_sean_snp2gff.pl -dir "../../../parameterize_assembly/test/melpo_cdna_assembly/chosen
/Melpomene_assembly_genbank.accurate.2_out.caf.named.ace_dir" -org H.melpomene -qual "../..
/../parameterize_assembly/test/melpo_cdna_assembly/chosen
/Melpomene_assembly_genbank.accurate.2_out.padded.fasta.qual.named" -orf ../../../prot4EST
/test/melpo_p4e/output/orf.all.fsa -link ../../../prot4EST/test/melpo_p4e/output
/prot_main.psql.named -pad 20 -table inv
```

For **annot8r** and **iprscan**, you will need to provide the physical properties outfile (.physprop) as a link and then you may specify one or more GO, EC, KEGG, iprscan outfiles. There is an extra experimental option for those expert users using chado: you can make use of the ontologies by specifying the -dsn option to a chado connection file (see perldoc). Otherwise, if using SeqFeature, specify -nodb. You can also specify the cut-offs (score and evalue) to be used. If you wish to use different one for KEGG, EC, GO etc then you should run them separately.

```
$ ic_annot8r2gff.pl -nodb -cs 80 -ce 1e-10 -phys ../../../annot8r/test/melpo_annot8r
/DM34740AaApep.phys  ../../../annot8r/test/melpo_annot8r/DM34740AaApep.fsa_1.in.iprscan  ../../..
/annot8r/test/melpo_annot8r/DM34740AaApep_vs_EC.blastp.annot.EC  ../../../annot8r
/test/melpo_annot8r/DM34740AaApep_vs_GO.blastp.annot.GO  ../../../annot8r/test/melpo_annot8r
/DM34740AaApep_vs_KEGG.blastp.annot.KEGG
```

General tips

Please note

1) Make sure you use the .named files**!** Without having standard ID, we cannot create GFFs.

2) for p4e you will notice we have many options. The -main file links the protein objects with the contigs. The -blast is not a BLAST report but the results of the BLAST parsing from prot4EST.

3) The -e option allows you to embed a copy of the FASTA sequences at the end of the GFF

4) Parsing uniref100 is quite slow due to all the HSP objects that BioPerl has to parse...

Populating a SeqFeature database for GBrowse

This step assumes you know enough about databases in order to use them confidently, so please note this is for advanced users.

The current version of est2assembly is aimed for users who wish to be able to analyze and display their transcriptomic data. It allows integration of GBrowse (and other applications) with Chado but we do not recommend it as it is very slow. In addition, the current version does not leverage the full data warehousing capabilities of Chado. Because we are building such a warehouse in our lab, do expect the next version to have this capability but please do note that chado has a steep learning curve and we don't recommend anyone but those versed in Bioinformatics to utilize it (e.g. if you don't know what a DAG is, you will have difficulties to populate chado).

The database schema of Bio::DB::SeqFeature is much easier to understand, use and much, much faster. It does not offer (or is meant to) the data warehousing capabilities of Chado but it is sufficient for the purpose of viewing data.

As mentioned previously, there are two types of GFFs, those with the biofeature tag in the filename and those without. The latter can be uploaded in chado with the following command

$ gmod_bulk_load_gff3.pl --noexon --dbname $DB_CHADO_NAME -dbhost $CHADO_DB_HOST -dbport $CHADO_DB_PORT --recreate_cache --analysis --organism "name of organism from organism table" -g gfffile

for SeqFeature things are easier using a script from the -dev branch of BioPerl:

See the help
$ bp_seqfeature_load.pl -h
or
$ perldoc bp_seqfeature_load.pl

I recommend using it with a postgres database (as mysql does not have the data integrity abilities of postgres). The command line can be quite tricky and it is tedious to load many GFFs. So I provide the script ic_loaddata.pl to help you.

$ perldoc ic_loaddata.pl

You can provide as many CAF- (assemblies), p4e- or BLAST-derived GFF files you want. For each CAF there must be a p4e file. The script will (re)create a SeqFeature database, making a backup if necessary. It will use your /tmp as a

temp directory to load the GFFs more quickly.

GBrowse integration:
In the gbrowse folder of this distribution there are some .inc files which you should customize for your system. If you copy them to your GBrowse 1.69 configuration folder (e.g. "/etc/apache/gbrowse.conf/") and pass the name of the directory to ic_loaddata.pl with the -config option, then you will create GBrowse 1.69 configuration files for each set of CAF/p4E file set. For each set you will have a _caf, a _orf and _prot .conf file, one for each page of GBrowse that shows the assembly, the ORF object and the protein object respectively.

So the command for our melpomene data (in one simple line) is:
From the gffcreation/test/melpo_gffs directory:

```
$ ic_loaddata.pl -caf Melpomene_assembly_genbank.accurate.2_out.caf.named.biofeature.gff -p
DM34740AaApep.fsa.biofeature -b
Melpomene_assembly_genbank.accurate.2_out.unpadded.fasta.x.uniref100.fasta.refblast.named.gff.biofeatu
DM34740AaApep.phys.gff DM34740AaApep_vs_EC.blastp.annot.EC.gff
DM34740AaApep_vs_GO.blastp.annot.GO.gff DM34740AaApep_vs_KEGG.blastp.annot.KEGG.gff
DM34740AaMsnp.gff DM34740AaApep.fsa_1.in.iprscan.gff -prefix dm_ -create -a "/etc/apache2
/gbrowse.conf/"
```

NB:

- You can use more than one -c -p -b -s -t option or you can specify many files one after the other in one -b etc option. This applies to any option which has the symbol {,} in the PerlDoc or usage instructions.
- If multiple assemblies are specified, however, you have to follow the same order in each of the -c -p -s.
- The -t option is optional, as it will be activated if you decide to load a p4e and/or caf file. You can use it to recreate the config files without loading new data.
- The -b option can handle any annotation: BLAST, SNPs, annot8r etc.
- the -a value might be different in your system.

So now if you use

```
$ psql -l
```

you should be able to see your database as dm_34740. I included a backup of this file using pg_dump :

```
$ pg_dump dm_34740 -CxOf dm_34740.psql
```

where -x and -O are used to remove personal info relating to my system, -f specifies the file and -C is used to include the creation commands in the .psql file so that you can restore the backup like this:

```
$ psql -f dm_34740.psql
```

That's it! Now you have a database and the GBrowse config files:

```
$ ls -trl "/etc/apache2/gbrowse.conf/"

-rw-r--r-- 1 alexie alexie    35979 2009-06-27 18:25 34740_prot.conf.bak
-rw-r--r-- 1 alexie alexie    35979 2009-06-27 18:25 34740_prot.conf
-rw-r--r-- 1 alexie alexie     4936 2009-06-27 18:25 34740_orf.conf
-rw-r--r-- 1 alexie alexie    28675 2009-06-27 18:25 34740_caf.conf.bak
-rw-r--r-- 1 alexie alexie    28675 2009-06-27 18:25 34740_caf.conf
```

Where .bak will exist if you run it more the once and it holds your backup. Each of these .conf files is scanned by GBrowse when it starts so it should be available under your Gbrowse installation as /34740_caf or /34740_prot or /34740_orf for the assembly, protein and ORF pages respectively.

The assembly _caf page links to the ORF _orf page which links to the protein _prot page. My demonstration is available here:
http://rfc.ex.ac.uk/cgi-bin/gbrowse/34740_caf/

When you click on it, you will get an empty GBrowse because no contig has been specified. But look at the tracks and compare with the prot and orf pages:
http://rfc.ex.ac.uk/cgi-bin/gbrowse/34740_prot/

http://rfc.ex.ac.uk/cgi-bin/gbrowse/34740_orf/

This manual is not a tutorial for Gbrowse, but very briefly:
What we need to do is specify a contig (e.g. DM34740AaEcon1), protein (e.g. DM34740AaApep515) or ORF (e.g. DM34740AaAorf515)to get a view. This can be given in the URL or in the search box ("Search Landmark or Region:")
http://rfc.ex.ac.uk/cgi-bin/gbrowse/34740_caf/?name=DM34740AaEcon1

http://rfc.ex.ac.uk/cgi-bin/gbrowse/34740_prot/?name=DM34740AaApep515

http://rfc.ex.ac.uk/cgi-bin/gbrowse/34740_orf/?name=DM34740AaAorf515

Let's look more closely how you interrogate the dataset using GBrowse in the next section.

Interrogating the dataset and understanding the structure

For this tutorial, you can use my demonstration website or the one you built for yourself using the test Melpomene data. Let's read also the GFF files (using your favourite editor/viewer) so that we understand how things are linked. First, the _caf assembly.

```
$ less Melpomene_assembly_genbank.accurate.2_out.caf.named.biofeature.gff
```

If you don't know what the GFF file format is, see here:
http://song.sourceforge.net/gff3.shtml (please note, the Sanger website is very much out of date).

Each line represents a feature which is anchored on the object specified in the left hand side column (the "reference"). Each feature has an ID, which needs to be unique in the whole database (i.e. across all .gff files loaded). This ID can be used in the Search box of GBrowse:

Try giving **EE743470** to the assembly page search box. It will automatically find the contig that contains this EST ID and you will get the EST shown and highlighted.

Now let's look at some annotation: Activate the **Known Proteins** track under **Annotation**. You will get the protein match that was included in your Uniref100 BLAST. Let's look at the GFF file:

```
$ less
Melpomene_assembly_genbank.accurate.2_out.unpadded.fasta.x.uniref100.fasta.refblast.named.gff.biofeatu
```

You will see the same hits in the file where the left hand side column (the reference) is the contig ID DM34740AaEcon2. Each line can have in the third column match_part or protein_match. The latter is the ueber-entity which groups the former, i.e. the HSP and HIT of the BLAST respectively. Let's pick one that has more than one HSP and let's search using the information contained in the annotation field (the last column on the right called 'attributes'). E.g. let's look at this line:

```
DM34740AaEcon3  BLASTX_uniref100        protein_match   268     417     4e-05   +       .
ID=DM34740AaEcon3:uniref100:Hit1;Alias=Similar to
UniRef100_UPI0000DA399F;Dbxref=uniref100:UniRef100_UPI0000DA399F;
Name=DM34740AaEcon3:uniref100:Hit1;Note=PREDICTED: hypothetical protein n%3D1 Tax%3DRattus
norvegicus RepID%3DUPI0000DA399F;algorithm=BLASTX;fraction_of_conserved_positions=0.0467;
fraction_of_identical_positions=0.0327;hit_rank=2;length=643;logical_length=227;
organism=H.melpomene;score=4e-05;top bit_score=52.0;top raw_score=123;total significance=4e-05
```

which has many match_parts (HSPs) following it. Look at the Note field and note the funny % characters: These are 'escape values' for URLs. They are not human-friendly but they are important for the safety of your server and will be translated in your browser. Now look at the Alias tag. It contains a protein ID from UniProt (**UniRef100_UPI0000DA399F**). The same ID is included in the Note field, with the UniRef100 ID stripped out (**3DUPI0000DA399F**). Copy paste **3DUPI0000DA399F** to your assembly search box and click search. You will get a list of all contigs that have this annotation. You can do this, with any part of the Note field, such as **Rattus norvegicus**. Now click to one of the search results and you will go to the contig view containing this hit.

Move your mouse over the hit ("hover") and you will get a description in a little balloon. Click and you will go to a detailed description which has a stable URL you can email to people:

http://rfc.ex.ac.uk/cgi-bin/gbrowse_details/34740_caf?name=DM34740AaEcon1%3Auniref100%3AHit1;class=Sequence;ref=DM34740AaEcon1;start=280;end=429;feature_id=7546

(Note the funny % characters, they are for your safety too). On this page you will get some links. They will point to the Uniref database for this protein. http://www.uniprot.org/uniref/?query=UniRef100_UPI0000DA399F&sort=score

Cool huh? I thought so too... (Thank you and well done Lincoln and Scott**!**).

So let's look at some Markers now. The only marker module included in est2assembly is SNPs (maybe others such as SSRs in another release). Here is the GFF we produced for it:

```
$ less DM34740AaMsnp.gff
```

You will note contig 107 (DM34740AaEcon107) has a lot of SNPs. So let's look at it...

http://rfc.ex.ac.uk/cgi-bin/gbrowse/34740_caf/?name=DM34740AaEcon107 and activate the **SNP track** under **Predictions**.

You will get Pie Charts of the various alleles and if you hover your mouse you will some more information. If you click you can see the details.

If you remember, we give a unique ID to each object, so our SNPs will also be unique and we can use it to search. So let's take **DM34740AaMsnp2** search for it. You will see that it is also aligning with the ORF so it must be coding. One way to access the ORF page is by clicking on the predicted protein shown in the assembly view.

NB: The SNP module is experimental, the main reason being that I find prot4EST (a.k.a. p4e) to be a great program but limited and buggy. I've been promised an excellent version 3 which will solve many problems so please be patient (as am I... but there is always the possibility I find time to reinvent the protein prediction wheel).

One nice thing about the way we set up our naming scheme and database is that we can have many protein (and thus ORF) predictions for one contig. You will notice that the name of the ORF, DM34740AaAorf337 is not similar to DM34740AaEcon107. The mapping was done during ic_create_naming.pl creating the link file prot_main.psql.named. This info is now stored in the GFF:

```
$ less DM34740AaApep.fsa.biofeature
```

You will notice that the protein predictions come in groups of four:

```
DM34740AaEcon59 placeholder     ORF     88      504     .       +       0
ID=DM34740AaAorf439_0;Dbxref=InsectaCentral:DM34740AaAorf439;Name=DM34740AaAorf439_0;conf_end=459;
conf_start=88;end_frame=1;ext_end=504;gene_location=nuclear;method=similarity;
organism=H.melpomene;start_frame=1
DM34740AaAorf439        prot4EST        ORF     1       417     .       +       0
ID=DM34740AaAorf439;Dbxref=InsectaCentral:DM34740AaAorf439;Name=DM34740AaAorf439;Note=Open
Reading Frame of protein prediction.;conf_end=459;conf_start=88;end_frame=1;ext_end=504;
gene_location=nuclear;method=similarity;organism=H.melpomene;start_frame=1
```

```
DM34740AaApep439        prot4EST        polypeptide     1       139     .       +       0
ID=DM34740AaApep439;Dbxref=InsectaCentral:DM34740AaApep439;Derives_from=DM34740AaEcon59;
Name=DM34740AaApep439;gene_location=nuclear;method=similarity;organism=H.melpomene
DM34740AaAorf439        placeholder     polypeptide     1       417     .       +       0
ID=DM34740AaApep439_0;Dbxref=InsectaCentral_DM34740AaApep439;Name=DM34740AaApep439_0;Note=HSP
similarity to uniref100 was 4e-50;gene_location=nuclear;method=similarity;organism=H.melpomene
```

The first feature is putting an ORF in respect to the assembly page (Econ is reference), the second feature initiates an ORF page (reference column is equal to the ID), the third does likewise for the protein object and finally the fourth links the protein object with the ORF object. We don't have to initiate the Econ (contig) object because this has been done in the .caf GFF file (take a look!).

So Econ59 has Aorf439 and therefore also Apep439 (Aorf and Apep serials also follow each other). If we wanted to add another protein prediction, it can be very easily done by finding out what is the next free unique ID for Aorf/Apep and giving it to the ic_create_naming.pl script when we process the p4e data. That way if for example we use different HMM for ESTScan (tier II of p4e) or database for similarity searches (tier I of p4e) then we can have both predictions sets appearing on the Gbrowse interface.

So let's look at a protein. Click on the protein track on the ORF page or just go to the protein URL:

http://rfc.ex.ac.uk/cgi-bin/gbrowse/34740_prot/?name=DM34740AaApep439

You will see that there are quite a few annotations here which do not appear in the contig page. This is because KEGG, GO, EC and InterProScan is run using proteins, not EST contigs. The main reason is that 6 frame translations of EST contigs (i.e. BLASTX) are notorious for being a) just terrible due to frameshifts decreasing the scores b) computationally 6 times more expensive than a BLASTP (which is not a big problem if you are interrogating 1000 contigs and have spare machines doing nothing but a huge problem when you're annotating hundreds of thousands or your computational resources are limited).

As usual, hovering your mouse and clicking gives more details for each annotation. These annotations are derived from the annot8r GFF so let's look at one of them (e.g. the GO).

```
$ less DM34740AaApep_vs_GO.blastp.annot.GO.gff
> /DM34740AaApep439
```

or to speed up the search by looking only at the beginnining of each line (what the little ^ denotes)

```
> /^DM34740AaApep439
```

Search for the protein object DM34740AaApep439 by using the search facility of your viewer/editor (for less it is forward slash /).

```
DM34740AaApep439        ANNOT8R_GO      supported_by_sequence_similarity        1       139
1e-28   .       .       ID=DM34740AaApep439:GO:3110;
```

```
Dbxref=DB:uniprot:Q9VZS5,local:5519,GO:0005515,GO:0005488;Name=DM34740AaApep439:GO:3110;
Note=protein binding (stat:0.03);organism=H.melpomene;score=1e-28
DM34740AaApep439        ANNOT8R_GO      supported_by_sequence_similarity        1       139
1e-40   .       .       ID=DM34740AaApep439:GO:3111;
Dbxref=DB:uniprot:Q5UAR0,local:4026,GO:0003735,GO:0005198;Name=DM34740AaApep439:GO:3111;
Note=structural constituent of ribosome (stat:0.90);organism=H.melpomene;score=1e-40
DM34740AaApep439        ANNOT8R_GO      supported_by_sequence_similarity        1       139
1e-40   .       .       ID=DM34740AaApep439:GO:3112;
Dbxref=DB:uniprot:Q5UAR0,local:5619,GO:0005622,GO:0005622;Name=DM34740AaApep439:GO:3112;
Note=intracellular (stat:0.90);organism=H.melpomene;score=1e-40
DM34740AaApep439        ANNOT8R_GO      supported_by_sequence_similarity        1       139
1e-28   .       .       ID=DM34740AaApep439:GO:3113;
Dbxref=DB:uniprot:Q9VZS5,local:6933,GO:0007052,GO:0009987;Name=DM34740AaApep439:GO:3113;
Note=mitotic spindle organization (stat:0.07);organism=H.melpomene;score=1e-28
DM34740AaApep439        ANNOT8R_GO      supported_by_sequence_similarity        1       139
1e-40   .       .       ID=DM34740AaApep439:GO:3114;
Dbxref=DB:uniprot:Q5UAR0,local:6346,GO:0006412,GO:0008152;Name=DM34740AaApep439:GO:3114;
Note=translation (stat:0.90);organism=H.melpomene;score=1e-40
DM34740AaApep439        ANNOT8R_GO      supported_by_sequence_similarity        1       139
1e-40   .       .       ID=DM34740AaApep439:GO:3115;
Dbxref=DB:uniprot:Q5UAR0,local:5817,GO:0005840,GO:0005622;Name=DM34740AaApep439:GO:3115;
Note=ribosome (stat:0.90);organism=H.melpomene;score=1e-40
DM34740AaApep439        ANNOT8R_GO      supported_by_sequence_similarity        1       139
1e-28   .       .       ID=DM34740AaApep439:GO:3116;
Dbxref=DB:uniprot:Q9VZS5,local:1823,GO:0000022,GO:0009987;Name=DM34740AaApep439:GO:3116;
Note=mitotic spindle elongation (stat:0.07);organism=H.melpomene;score=1e-28
DM34740AaApep439        ANNOT8R_GO      supported_by_sequence_similarity        1       139
7e-39   .       .       ID=DM34740AaApep439:GO:3117;
Dbxref=DB:uniprot:Q962T2,local:15536,GO:0030529,GO:0005622;Name=DM34740AaApep439:GO:3117;
Note=ribonucleoprotein complex (stat:0.07);organism=H.melpomene;score=7e-39
```

You will notice that "protein binding" is in the Notes. You can use anything in the Note field to search your GBrowse. Try putting **intracellular** on the search box of the protein page.

This is pretty much all you need to know to use the database. If something is amiss, drop me a line at alexie -alpha symbol- butterflybase.org and I'll get back to you as soon as possible.

Happy gene hunting**!**

Using a LSF-driven PC-Farm

If you are doing many transcriptome projects, then computational power will be the limit, especially for annotation. It really helps if you have access to a PC-farm or a computing cluster. If you have a job management system based on LSF then you use the scripts I use for annotation. If you do have access to a computing cluster, I assume (and trust) that you fall in the **expert** user category so I'll keep this brief. Please note that the scripts are provided for the sake of your convenience and you will probably need to edit them a bit to customize it to your system. Hopefully will not spend as much time as if you wrote your own.

Filesystem Structure

We assume there is an NSF directory (e.g. your home) shared across all nodes. We also assume that each node has a scratch disk at /tmp that you can write to (about a couple of gigabytes will be needed, depending on what you're trying to BLAST).

You will need to copy all the files in the **lsf** directory to your home directory. The directory structure should be maintained as is but can be linked to any FileSystem you wish.
For example, we use a non-backed up FS for transfer, storage and tmp.

Databases should reside bzipped in the dbs/bz2 directory and a md5sum file at dbs/
you can use this command to create it from within "~/dbs/bz2"
$ md5sum *bz2 >../md5sums.txt

The FASTA directory contains some FASTA files which or may not be of use to your LSF approach (e.g. run SSAHA on a cluster)

BLASTs

The prepare_lsf.pl is used to prepare for blast runs. Run it from the "~/blasts /blast_in/" See the run_annot_afterparameterizing.sh for a bash file that will automatically generate all the required directories, files and launch the jobs if run from the above directory. The inputs of this file are just fasta files (also residing at "~/blasts/blast_in/".

Once the blasts are complete (or even before) you can use check_blasts_lsf.pl to see which have finished and relaunch them. There are some hard-coded values in this script to protect me from launching from execute nodes instead of the submit nodes (deimos): you will have to change them.

Because the scripts rely on the nodes having a /tmp directory where you can copy the databases (it really improves performance a lot...), one clever thing you can do is copy the databases to the nodes beforehand. Your system administrator will also love you as you will not have 100s of open files across the NSF.

InterProScan (IPRSCAN)

For this program you can use the run_iprscan.pl program. The iprscan has to be properly installed beforehand, including all the /data libraries that are shipped from EBI.
run_iprscan.pl is used like like prepare_lsf.pl but takes no options: the only arguments is a list of FASTA files (from within "~/blasts/blast_in/").

est2assembly citation

Alexie Papanicolaou, Remo Stierli, David G. Heckel and Richard ffrench-Constant: Next generation transcriptomes for next generation genomes using ***est2assembly***. In preparation: please ask for an update

Remember: you **must** cite **all** of the programs used in the pipeline (see INSTALL file) and also Chado and GMOD. For software, it is an omission which is not uncommon but it is not very nice (and is academic misconduct). Typically, one includes in their methods: Data processed by [platform - est2assembly in this case] [ref] which uses prog1 [ref], prog2 [ref] etc. If you choose not to cite them properly, then the publication citation indexes of the software is not increased: i.e we get no public funding to provide the methods you just used for free to publish your data...

Acknowledgements

From our paper:
We would like to thank the following for making pre-publication data available: Chris Jiggins and his laboratory, Owen McMillan and his laboratory, Yannick Pauchet, Iva Fuková, Haobo Jiang and Heiko Vogel. Bastien Chevreux provided development versions of MIRA and excellent support, Jose Blanca provided sff_extract, James Wasmuth provided support for prot4EST, Ralf Schmid for annot8r, Derek Huntley for SEAN. David Clements and Scott Cain helped with chado and Gbrowse. We also thank the TU-Dresden Deimos PC-Farm for computational support. The authors report no conflicting interests. AP conceived, designed and performed the study; analysed and interpretated data; coded the software and drafted the manuscript. RS co-authored the GFF writing software and the Gbrowse schema. RHfC drafted the manuscript, financed and provided infrastructure for the study. All authors approved the final version of the manuscript. AP was supported by a Max Planck Stipendium by the Department of Entomology (MPI for Chemical Ecology) and the European Union Research Network GAMEXP.

DISCLAIMER & LICENSE

Appendix B – Genes differentially expressed in the *Manduca sexta* dataset

Data presented here are without sequences due to confidentiality with the owner of the data (Dr Yannick Pauchet). The appendix containing only contig IDs is published with Dr Pauchet's consent.

Contig	C1_corrected	C2_corrected	T1_corrected	T2_corrected	T4_corrected	P	Q	FDR genes	False genes	lfdr	T_means	C_means
IC7130AgEcon5779	57.73	53.12	2738.9	2418.44	2341.27	0	0	3	0	0	2499.54	55.43
IC7130AgEcon6128	576.56	534.13	1699.32	1603.3	1675.85	0	0	4	0	0	1659.49	555.35
IC7130AgEcon9430	139.13	150.88	484.34	519.95	468.9	0	0	9	0	0	491.06	145.01
IC7130AgEcon514	4946.4	5045.38	8692.64	8485.67	9348.99	0	0	10	0	0	8842.43	4995.89
IC7130AgEcon4524	637.46	667.06	2574.63	3086.49	2672.73	0	0	12	0	0	2777.95	652.26
IC7130AgEcon7660	299.87	326.4	904.43	1009.18	1014.77	0	0	20	0	0	976.13	313.14
IC7130AgEcon1633	22605.21	26683.5	65085.96	63311.42	59393.17	0	0	22	0	0	62596.85	24644.36
IC7130AgEcon368	3045.09	3250.56	8058.23	6808.06	8016.11	0	0	27	0	0	7627.47	3147.83
IC7130AgEcon11504	187.32	183.23	525.97	457.21	505.67	0	0	31	0	0	496.28	185.28
IC7130AgEcon6203	203.45	151.08	1040.68	861.5	806.33	0	0	33	0	0	902.84	177.27
IC7130AgEcon6034	565.06	663.43	1298.94	1397.85	1478.03	0	0	38	0	0	1391.61	614.25
IC7130AgEcon1399	298.67	287.61	636.31	693.08	728.07	0	0	40	0	0	685.82	293.14
IC7130AgEcon13678	1584.66	1940.36	5127.54	4449.7	4896.61	0	0	41	0	0	4824.62	1762.51
IC7130AgEcon5184	264.3	329.02	1006.09	975.66	1235.76	0	0	42	0	0	1072.5	296.66
IC7130AgEcon2158	49.57	60.33	193.18	174.5	166.82	0	0	45	0	0	178.17	54.95
IC7130AgEcon10512	25.52	40.67	230.81	190.46	180.55	0	0	46	0	0	200.61	33.1
IC7130AgEcon16165	33.56	48.11	308.39	326.51	235.9	0	0	49	0	0	290.27	40.84
IC7130AgEcon1463	2203.55	2483.44	7330.58	10006.88	8641.81	0	0	54	0	0	8659.76	2343.5
IC7130AgEcon1777	414.72	557.99	2020.29	1778.2	1700.47	0	0	55	0	0	1832.99	486.36
IC7130AgEcon195	1737.29	1722.12	3107.52	3715.87	3294.09	0	0	60	0	0	3372.49	1729.71
IC7130AgEcon19683	3244.93	3980.19	11103.87	14784.85	12535.88	0	0	62	0	0	12808.2	3612.56
IC7130AgEcon3421	126.46	141.36	309.03	346.33	289.63	0	0	67	0	0	315	133.91
IC7130AgEcon2640	2830.44	3543.5	10551.76	9204.82	12409.82	0	0	72	0	0	10722.13	3186.97
IC7130AgEcon6196	148.3	100.34	486	471.33	403.05	0	0	75	0	0	453.46	124.32
IC7130AgEcon1415	**1544.4**	**1615.2**	**2941.91**	**2679.16**	**3116.13**	**0**	**0**	**79**	**0**	**0**	**2912.4**	**1579.8**
IC7130AgEcon4411	349.3	500.42	1812.74	2437.53	2023.94	0	0	80	0	0	2091.4	424.86
IC7130AgEcon4113	28.98	34.81	235.49	319.73	191.49	0	0	82	0	0	248.9	31.9
IC7130AgEcon13827	212.69	253.62	1172.44	1837.15	1952.92	0	0	83	0	0	1654.17	233.16
IC7130AgEcon14579	461.68	461.02	808.33	868.2	990.58	0	0	84	0	0	889.04	461.35
IC7130AgEcon489	687.95	804.25	2497.87	2596.72	1909.28	0	0	89	0	0	2334.62	746.1
IC7130AgEcon14398	164.35	126.1	426.71	471.53	372.45	0	0	92	0	0	423.56	145.23
IC7130AgEcon9985	100.17	90.43	1017.58	870.79	553.6	0	0	94	0	0	813.99	95.3
IC7130AgEcon18	5202.56	7581.48	23253.72	21212.63	21247.23	0	0	97	0	0	21904.53	6392.02
IC7130AgEcon3358	294.12	387.55	823.58	901.75	988.66	0	0	102	0	0	904.66	340.84
IC7130AgEcon11863	38.11	60.38	182.7	158.49	157.59	0	0	104	0	0	166.26	49.25
IC7130AgEcon255	2553.66	2687.49	5703.68	6099.78	4809.51	0	0	107	0	0	5537.66	2620.58

IC7130AgEcon170	1461.88	1559.37	9909.18	13660.25	7387.65	0	0	110	0	0	10319.03	1510.63
IC7130AgEcon6974	78.22	66.08	217.24	168.29	173.03	0	0	114	0	0	186.19	72.15
IC7130AgEcon1731	86.29	94.39	215.33	262.75	292.56	0	0	116	0	0	256.88	90.34
IC7130AgEcon7398	66.81	101.05	259.44	299.95	255.76	0	0	117	0	0	271.72	83.93
IC7130AgEcon3275	78.23	77.63	209.86	166.8	214.38	0	0	122	0	0	197.01	77.93
IC7130AgEcon1546	6584.88	4850.04	16792.26	19335.72	15509.87	0	0	130	0	0	17212.62	5717.46
IC7130AgEcon4374	64.44	55.92	143.35	177.96	142.37	0	0	136	0	0	154.56	60.18
IC7130AgEcon2560	157.44	131.49	313.77	400.72	382.95	0	0	137	0	0	365.81	144.47
IC7130AgEcon10771	205.67	178.37	481.12	535.51	398.63	0	0	139	0	0	471.75	192.02
IC7130AgEcon2582	581.17	553.54	1462.13	2278.64	1895.14	0	0	141	0	0	1878.64	567.36
IC7130AgEcon4853	384.76	467.35	736.34	752.6	748.18	0	0	142	0	0	745.71	426.06
IC7130AgEcon788	58.84	74.11	645.36	716.61	369.63	0	0	143	0	0	577.2	66.48
IC7130AgEcon453	2099.03	2636.17	5018.55	6528.7	6404.16	0	0	145	0	0	5983.8	2367.6
IC7130AgEcon14196	77.07	87.51	156.92	180.97	183.63	0	0	146	0	0	173.84	82.29
IC7130AgEcon5811	65.64	95.85	221.05	193.96	224.96	0	0	149	0	0	213.32	80.75
IC7130AgEcon8499	89.69	99.01	185.76	173.2	217.55	0	0	151	0	0	192.17	94.35
IC7130AgEcon1454	256.26	376.42	1166.54	962.77	909.18	0	0	153	0	0	1012.83	316.34
IC7130AgEcon2103	776.3	884.97	1696.72	1762.37	1403.23	0	0	154	0	0	1620.77	830.64
IC7130AgEcon1490	10043.16	11705.3	30003.42	21965.93	28282.2	0	0	160	0	0	26750.52	10874.23
IC7130AgEcon432	10048.9	11725.19	30036.82	21985.19	28302.14	0	0	161	0	0	26774.72	10887.05
IC7130AgEcon903	1062.31	1197.85	5009.78	8930.66	5301.66	0	0	163	0	0	6414.03	1130.08
IC7130AgEcon15955	191.93	205.76	478.62	548.63	703.31	0	0	164	0	0	576.85	198.85
IC7130AgEcon221	2785.61	3692.74	7461.9	6777.48	6875.2	0	0	165	0	0	7038.19	3239.18
IC7130AgEcon647	4029.01	4486.32	6516.67	6812.86	7063.88	0	0	166	0	0	6797.8	4257.67
IC7130AgEcon9420	202.22	228.97	382.57	453.91	488.76	0	0	171	0	0	441.75	215.6
IC7130AgEcon13413	328.46	274.03	538.7	503.84	502.84	0	0	172	0	0	515.13	301.25
IC7130AgEcon1195	73.64	86.14	172.16	211.48	166.96	0	0	174	0	0	183.53	79.89
IC7130AgEcon685	2333.17	2039.03	4456.15	4216.88	3538.13	0	0	179	0	0	4070.39	2186.1
IC7130AgEcon309	7045.31	9139.37	16993.84	19011.79	21765.94	0	0	180	0	0	19257.19	8092.34
IC7130AgEcon6853	156.39	204.3	829.33	566.62	945.72	0	0	181	0	0	780.56	180.35
IC7130AgEcon240	3000.29	3738.35	6387.89	6311.9	5944.35	0	0	184	0	0	6214.71	3369.32
IC7130AgEcon1727	623.59	509.88	1592.79	1308	1162.39	0	0	186	0	0	1354.39	566.74
IC7130AgEcon7885	2828.07	3690.19	6159.81	6071.11	5749.57	0	0	187	0	0	5993.5	3259.13
IC7130AgEcon10324	171.27	251.92	503.49	554.92	524.14	0	0	191	0	0	527.52	211.6
IC7130AgEcon1166	5529.71	4342.65	12455.54	16620.92	11743.23	0	0	202	0	0	13606.56	4936.18
IC7130AgEcon6749	515.62	532.76	898.83	829.74	791.25	0	0	203	0	0	839.94	524.19
IC7130AgEcon973	144.85	194.44	425.11	450.51	361.77	0	0	206	0	0	412.46	169.65

IC7130AgEcon6105	2784.44	4180.53	5081.74	6420.97	6056.23	0	0.01	937	9.03	0.05	5852.98	3482.49
IC7130AgEcon384	1529.52	1761.41	2370.31	4927.98	4191.04	0	0.01	940	9.34	0.06	3829.78	1645.47

Curriculum vitae

Alexie Papanicolaou

(formally: Αλέξιος Παπανικολάου, Alexios Papanikolaou)

Postal addresses: 72 Longfield, Falmouth; TR11 4SL; UK
CSIRO Ento; Black Mountain Labs; Clunies Ross; Acton 2601; Australia

Nationality: Hellenic (Greek) ; DOB: 14-July-1981 ; Sex: Male

Current employment:

- CSIRO CSE - Entomology; Office of the Chief Executive Post-Doctoral Fellow in Bioinformatics (since July 2010)

alexie@butterflybase.org

Education / Academic Employment	
Research associate: University of Exeter in Cornwall; CEC-Biology; Tremough Campus; TR10 9EZl UK	*Jan 2009 – May 2010*
Ph.D candidate: Emerging model species driven by transcriptomics Max Planck Institute for Chemical Ecology, Jena, Germany (Jan 2006 -) Due submission date: 1st March 2010	*Jan 2006 - present*
M.Sc. by research - Evolutionary Genetics – Institute of Evolutionary Biology, University of Edinburgh, UK (Distinction)	*Sep 04 ~ Sep05*
B.Sc. Genetics (Hons) – University of Edinburgh, UK (IIi). 2003 thesis on miRNA comparative genomics in Drosophila sp.	*Sep 00 ~ Jun 03*
Bench Supervisor of PhD candidate Ritika Chauhan (U. Exeter) Bsc Hons Victoria Renders Computer Science M.Sc. student Remo Stierli (U. Rhode Island). Undergraduate assistants Charles Imbusch and Saskia Wolfram	 *Sep 2009 – May 2010* *Sep 2009 – June 2010* *2008 to 2009* *2006 to 2007*
Teaching assistant. Evolutionary Biology module	*2005~2006*
Graduate research assistant. Dr. Chris Jiggins & Prof. M. Blaxter. Molecular marker development *Heliconius species*	*April 04 ~Sep04*
Graduate research assistant. Dr. Patricia Lee, University of Swansea. Molecular marker development for *Brassicaceae sp*.	*Jan 04 ~ Apr04*
Undergraduate and graduate research assistant. Dr. Peter Andolfatto, University of Edinburgh/Toronto. *Drosophila* genomics	*Jan 03 ~Jun03 Aug 03 - Sep03*
Database **Technical Assistant**. Drs. P. Preston, B.E. Matthews (Edinburgh). *Ixodiae* collection.	*Jun 01 ~ Aug 01*
Zoology research team Greek Ornithological Society, Gialova Lagoon, Pylos, Greece (EU Natura 2000 funded).	*1999 ~ 2002*

Publications

1. Papanicolaou A., Stierli R., ffrench-Constant H.R., Heckel DG (2009) Next generation transcriptomes for next generation genomes using est2assembly. BMC Bioinformatics 10:447 (Highly Accessed)
2. Ferguson L, Lee SF, Chamberlain N, Nadeau N, Joron M, Baxter S, Wilkinson P,

Papanicolaou A, Kumar S, Kee TJ, Clark R, Davidson C, Glithero R, Beasley H, Vogel H, ffrench-Constant RH, Jiggins CD (2010) Characterization of a hotspot for mimicry: Assembly of a butterfly wing transcriptome to genomic sequence at the HmYb/Sb locus. Molecular Ecology special issue

3. Papanicolaou, A. and Gebauer-Jung, S. and Blaxter, M.L. and Owen McMillan, W. and Jiggins, C.D. (2008) ButterflyBase: a platform for lepidopteran genomics. Nucleic Acids Research 36:D582 (Faculty 1000: f1000biology.com/article/id/1097096)
4. Beldade P; McMillan WO; Papanicolaou A; (2007): Evolutionary & Ecological Functional Genomics special issue, commissioned review on butterflies: Butterfly Genomics Eclosing Heredity 98
5. Pringle, B., Baxter, S. W., Webster, C. L., Papanicolaou, A., and Jiggins, C. D. (2007) Synteny and chromosome evolution in the Lepidoptera: Evidence from mapping in Heliconius melpomene Genetics 177(1):417
6. Joron, M; Jiggins, CD; Papanicolaou A; McMillan WO (2006): Evolutionary & Developmental Biology special issue, commissioned review: Heliconius wing patterns: an evo-devo model for understanding phenotypic diversity. Heredity 97(3):157-67
7. Papanicolaou A; Joron M; McMillan WO; Blaxter ML; Jiggins CD (2005): Genomic tools and cDNA derived markers for butterflies. Molecular Ecology 14:2883-2897

Manuscripts @ submission

Papanicolaou A. Heckel DH. The Drupal Bioinformatic Server Framework (Bioinformatics) – under revision

Papanicolaou A, Heckel DH. InsectaCentral: Transcriptomes for molecular ecology & evolutionary and ecological functional genomics (in prep)

H publication index = 5

Talks

Papanicolaou A. InsectaCentral – Facilitating transcriptome work through a GMOD platform. 1st International Workshop on Information Systems for Insect Pests, Rennes France 16-17th November 2009

Pauchet Y, Papanicolaou A. Vogel H, Heckel DG, ffrench-Constant R. *Manduca* midgut

transcriptomics; The 8th International Workshop on Molecular Biology and Genetics of the Lepidoptera / Orthodox Academy of Crete, Kolympari, Crete, Greece, Aug 2009

Papanicolaou A: The wide utility of Lepidopteran genomic resources; The Seventh International Workshop on Molecular Biology and Genetics of the Lepidoptera / Orthodox Academy of Crete, Kolympari, Crete, Greece, Aug 2006

Papanicolaou Alexie: Genomic tools for Lepidoptera; 4th Biannual IMPRS Symposium / MPI for Chemical Ecology, Jena, Germany, Mar 2006.

Papanicolaou Alexie: Genomic tools for butterflies and applications to speciation research; ICE Department Seminar, 2005 Apr 26

Posters

Papanicolaou Alexie, Chris D. Jiggins, Owen W. McMillan, David G. Heckel. ButterflyBase: A framework for comparative genomics in butterflies and moths SMBE meeting June 2008

Papanicolaou Alexie, Chris D. Jiggins, Owen W. McMillan, David G. Heckel. ButterflyBase: A framework for comparative genomics in butterflies and moths. Arthropod Genomics meeting April 2008

Papanicolaou Alexie, Chris D. Jiggins, Owen W. McMillan, David G. Heckel. ButterflyBase: an organismal database and resource for Lepidoptera. Biocurators Meeting, San Jose Nov. 2007

Papanicolaou Alexie, Heckel David, Schöfl Gerhard, Wolfram Saskia: Incipient speciation in Spodoptera frugiperda: Dissecting with many scalpels. SAB Meeting 2006 / MPI for Chemical Ecology, Jena, Germany, Oct 2006

Papanicolaou Alexie, Schöfl Gerhard: Emergence of host races in Spodoptera frugiperda: reproductive isolation; ICE Symposium / MPI for Chemical Ecology, Jena, Germany, Jun 2006

Other

Co-organizer of Workshop at Crete Aug 2009: Lep genome annotation and bioinformatics solutions, coordinated by Goldsmith M., Papanicolaou A., Legeai F.).

Peer-reviewer for “Insect Molecular Biology”, “Genome”, “Computers and Electronics in Agriculture” and Nucleic Acids Research.

Designer and curator of ButterflyBase, the annotated EST database of Lepidoptera. Designer and curator of InsectaCentral, a GMOD platform for gene curation

Acknowledgements

This work would not have been possible if a number of senior researchers had not supported me by providing the necessary finances, data and intellectual support: David G. Heckel as my doctoral supervisor, Richard ffrench-Constant, Karl H.J. Gordon and especially Dave Clements & W. Owen McMillan who kept me inspired. Earlier work was based on collaborations and conversations with Mark L. Blaxter and Chris D. Jiggins to whom I'm grateful for supporting my early career. Data provided by Heiko Vogel, Yannick Pauchet, Vic Renders, Lars Jermiin and Iva Fuková were essential in making this software functional in the real world. Conversations with a countless people have contributed to concepts developed and presented here but I could not begin naming them in fear of forgetting someone. Most important of all, none of this would have started without the loving support of my parents, Mimi and Antonis Papanicolaou, and Iva Fuková who kept me sane during the trials of a PhD thesis.

Selbständigkeitserklärung

Die geltende Promotionsordnung der Biologisch-Pharmazeutischen Fakultät der Friedrich Schiller-Universität ist mir bekannt. Die vorliegende Dissertation habe ich selbständig verfasst und keine anderen als die von mir angegebenen Quellen, persönliche Mitteilungen und Hilfsmittel benutzt. Es wurden keine Textabschnitte eines Dritten ohne Kennzeichnung übernommen. Alle Personen, die an der Gewinnung von Daten beteiligt, bei der Erstellung des Manuskripts hilfreich waren oder sonstige Hilfestellungen gaben, sind benannt. Es wurde weder bezahlte noch unbezahlte Hilfe eines Promotionsberaters in Anspruch genommen. Ich habe die Dissertation noch nicht als Prüfungsarbeit für eine staatliche oder andere Wissenschaftliche Prüfung eingereicht.

Canberra, Australia, den 06. December 2010

Alexie Papanicolaou